Comparing Validity Evidence of Two ECERS-R Scoring Systems

Songtian Zeng

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Susan R. Sandall, Chair

Gail Joseph

Kathleen Artman Meeker

Min Li

Program Authorized to Offer Degree:

College of Education

University of Washington

Abstract

Comparing Validity Evidence of Two ECERS-R Scoring Systems

Songtian Zeng

Chair of the Supervisory Committee:

Professor Susan R. Sandall

Special Education

Over 30 states have adopted the Early Childhood Environmental Rating Scale-Revised (ECERS-R) as a component of their program quality assessment systems, but the use of ECERS-R on such a large scale has raised important questions about implementation. One of the most pressing question centers upon decisions users must make between two scoring systems: stop scoring, in which scoring on an item is ceased if indicators of lower categories are not fulfilled, and alternative scoring, in which all indicators are scored regardless of anchor scores. This question has implications not just for researchers interested in the psychometric properties of assessments, but also for coaches who use the ECERS-R for training, coaching, or technical assistance in their state's QRIS. The purpose of this study, therefore, was to compare the differences of validity evidence based on the two scoring systems in the context of a state's QRIS. Utilizing a state representative early childcare sample collected in 2013-2015, I evaluated the descriptive differences between the two scoring methods and compared the convergent validity with the CLASS tool. Moreover, I conducted a series of regressions to examine its predictive validity with child learning outcomes. To gather consequential validity data, I interviewed 13 coaches about their use of ECERS-R scoring systems to identify coaching goals and to implement data-

based decision making. Quantitative findings suggested that the quality scores between the two scoring systems could be dramatically different. Some ECERS-R subscales significantly predicted children's language and early math outcomes. However, the effect sizes were small for both scoring systems. Qualitative findings indicated that coaches felt frustrated with expectations with some quality indicators due to lack of feasibility, cultural adaptation, and compatibility with program philosophy and facilities. Coaches preferred the alternative scoring system as it provided more information and cultivated a strength-based coaching partnership. Data-based decision making was grounded in valid information, coaches' knowledge about the tools and the programs, and the ability to interpret and negotiate with program providers, which could vary across coaches. This study is the first attempt to deepen our understanding of how scoring systems may affect other aspects of validity evidence. Implications and suggestions for future research direction are provided.

Keywords: ECERS-R, validity, scoring system, data-based decision making

ACKNOWLEDGEMENTS

I want to express my sincere gratitude to my dissertation committee for their mentoring in the past years. My advisor, Dr. Susan R. Sandall has been very supportive and provided me with many great learning opportunities over the past three years. Although Susan is retiring, her ideas about inclusion, high-quality learning opportunities, teacher preparation, family support, and program quality improvement will continue to benefit the field. I would like to thank Dr. Gail Joseph, who gave me invaluable working experience at the center of Childcare Quality and Early Learning Center for Research and Professional Development. I want to thank Dr. Kathleen Artman Meeker. Kathleen is a great role model and I admire her as a productive scholar. I am truly grateful for Dr. Min Li's guidance and statistical support in conducting this project. I thank for Dr. Jason Yip's valuable insights and thoughtful comments in this manuscript. I am grateful to receive professional guidance from my committee, and I appreciate their encouragement throughout my doctoral study. I want to thank my writing group which provided me a "safe space" for critical feedback. They are Dr. Janet Soderberg, Dr. Robert Abbott, Dr. Pooja Tandon, Dr. Sara Stull, Dr. Katherine Lewis, Nail Hassairi, Linghui Chu, Sayaka Omori, Lindsey Wilson, Xueyan Yang, Carlyn Mueller, Somin Yeon, and Jocelyn Walsh. Thanks Bamford Foundation for their generous financial support of my research work. I would like to express my gratitude to colleagues at the Center for Strong Schools (Dr. Gregory Benner, Kelcey Schmitz, Rayann Silva, Leigh Butler, and Lauren Ashbaugh) for their encouragement and support in the past two years. I highly appreciate the participants in this study who help advance our thinking about early childhood program quality assessment and improvement. Finally, I want to thank my wife Jane Li, and my family for their unconditional support.

TABLE OF CONTENTS

Chapter One: Introduction

High-quality early childhood education (ECE) experiences can have a profound influence on young children's school learning and long-term outcomes, particularly for children experiencing poverty or family hardships (Schweinhart et al, 2005; Yoshikawa et al., 2013). However, research demonstrates that many ECE programs do not meet established benchmarks for the provision of high quality early learning experiences (Adams, Tout & Zaslow, 2007). With limited resources, funders and policymakers are particularly interested in identifying active ingredients of high-quality ECE programs and investing strategically in initiatives designed to improve program quality, and thereby outcomes for young children (Forry, Vick, & Halle, 2009; Layzer & Goodson, 2006; Zaslow et al., 2006; Zaslow, Martinez-Beck, Tout, & Halle, 2011). Such initiatives are a core reason for the proliferation of Quality Rating and Improvement Systems (QRIS) across states (Boller, Tarrant, & Schaack, 2014). Typically, a state QRIS includes two components: (a) a quality assessment system and (b) resources (e.g., coaching; technical assistance; funding) that can support quality improvement (Child Trends, 2014; Zellman & Karoly, 2012).

Validating quality assessment system is critical for the success of quality improvement initiatives. Measurement tools can help coaches identify areas of strengths and weakness, set goals for improvement, allocate financial resources toward the most pressing needs and monitor progress over time (Zaslow, Tout, Martinez-Beck, 2010). A number of states, as well as federal programs such as Head Start, have used quality measurement for these diagnostic and descriptive purposes (U.S. Department of Health and Human Services, 2003; Barnard, Smith, Fiene & Swanson, 2006; Tout & Sherman, 2005).

As a force for change, QRIS heavily relies on high-quality measurement tools that can capture the "active ingredients" of program quality. Research has demonstrated associations between various aspects of classroom quality and improved social and academic outcomes for young children (Belsky et al., 2007; Helburn et al., 1995; Howes et al., 2008; Mashburn et al., 2008). However, defining these constructs and measuring them continue to challenge the field both methodologically in research and in quality enhancement initiatives (La Paro, Thomason, Lower, Kintner-Duffy, & Cassidy, 2012).

Defining Quality Constructs

Historically, definitions of quality in ECE included multiple proximal (e.g., curriculum and classroom interactions) and distal (e.g., program and state policies) features of classrooms that are believed to promote children's development in various domains (Dunn, 1993). However, the definitions of quality are often very broad and diverse due to disagreement among researchers and minimal validity evidence (Layzer & Goodson, 2006). La Paro et al (2012) systematically reviewed 76 studies published from 2003 and 2010 and found that a wide variety of definitions are used to conceptualize and operationalize ECE quality.

In recent years, a consensus has emerged on defining ECE global quality with two primary constructs— *structural* and *process* qualities (Vandell & Wolfe, 2000). Structural quality refers to easily measurable characteristics such as child: adult ratio, group class size, caregiver formal education, and caregiver specialized training related to children. Structural quality is conceptualized as the physical aspects of child care environment. Meanwhile, process quality is linked to the dynamic experiences that children have throughout the day, including human interactions occurring in the classrooms such as teacher-child and peer-to-peer interactions

(Cassidy et al., 2005; Hamre & Pianta, 2008; Phillipsen et al. 1997). Several instruments have been developed with indicators capturing the structure and process constructs.

ECERS-R and Program Quality Constructs

Early Childhood Environmental Rating Scale (ECERS), and subsequently the ECERS-R (Harms, Clifford, & Cryer, 2005), was developed to evaluate the process quality in settings for children. ECERS-R is used in more than 30 states' QRIS (alone or in combination with other measures; Child Trends, 2014). The ECERS-R consists of 43 items organized into seven subscales that guide the observer to relevant areas in early childhood classrooms. These include aspects of space and furnishings, personal care routines, language-reasoning, activities, interaction, program structure, and parents and staff. Recently the third edition ECERS-3 (Harms et al., 2015) was published. It encompasses changes consisting of refining some indicators and removing the items scored through teacher interview (in particular the sub-scale Parents and Staff). However, it retains much of the structure of the earlier versions and many states still use the ECERS-R in their QRIS systems.

The use of ECERS-R on such a large scale has raised important questions about implementation. One of the most pressing questions centers upon decisions users must make between two scoring systems: stop scoring, in which scoring on an item is ceased when an individual indicator is scored No, and alternative scoring, in which all indicators are scored regardless of anchor scores. This question has implications not just for researchers interested in the psychometric properties of assessments, but also for coaches who use the ECERS-R for training, coaching, or technical assistance in their state's QRIS. Next, I describe the differences between the two scoring systems.

Two Scoring Systems

ECERS-R captures program quality features in the qualitative indicators. Each item is described by several qualitative indicators for the odd-numbered response categories (1 indicating inadequate quality, 3 indicating minimal quality, 5 indicating good quality, and 7 indicating excellent quality). The qualitative indicators are supposed to assess the extent to which children's basic needs (protection of their health and safety, the facilitation of building positive relationships, and opportunities for stimulation and learning from experience) are met in a certain context (Harms et al., 2005).

ECERS-R items may be scored using the stop or alternative rules. The ECERS-R *stop-scoring* system postulates a specific order in item scores, since indicators of higher categories can only be rated if indicators of lower categories are fulfilled. Most of the items are structured so that indicators regarding health and safety or the availability of materials, space, and furnishings are placed in the lower categories while quality of interaction and instruction appear in the mid and high categories of the item. To receive high scores on one item, all the qualitative indicators of the lower categories need to be met (Harms et al., 2005).

Meanwhile, the developers propose an *alternative scoring* system. Specifically, the assessors are required to score all the indicators beyond the quality level score assigned to an item as each indicator can be given a score. Using the alternative scoring system can provide additional information on areas of strength beyond the quality level score and may be helpful in making plans for specific improvements and interpretation of research findings. In ECERS-3, the developers highly recommend scoring all indicators for all items when using the scale, but still use the stop-scoring system to measure program quality (Harms et al., 2015).

The use of two scoring systems (i.e., stop and alternative scoring) may have important implications in research and practice. Scoring is the foundational aspect of validity. Other

advanced analyses (e.g., predictive validity, factor analysis) are based on the assumption that scoring rule is appropriate. Different scoring systems may also influence the outcome interpretation and social validity. Preliminary research (Hofer, 2008, 2010) indicates that 25% of the childcare centers can move above one state's cutoff for more funding using alternative scoring instead of stop scoring methods. Yet, what remains unknown is whether the alternative scoring system may actually better reflect the factor structure and thus improve predictive validity of child outcomes. The current scoring recommendation is not based on empirical evidence (Gordon et al, 2015). No research, however, systematically compares the difference of validity inferences based on these two scoring methods. It is important to investigate this issue because validity claims about program quality factors based on inappropriate scoring system may not only jeopardize the quality of childcare research, but could also provide misleading information for coaches and policymakers.

To fill this high-stake knowledge gap and inform childcare quality improvement research, practices, and policy making, I conceptualize a mixed method study to exam the validity of ECERS-R. The purpose of this study is to compare the validity evidence of ECERS-R based on two scoring systems. The research questions include:

1. What were the score differences between the two scoring methods?
2. What was the convergent validity between ECERS-R and another quality measure (i.e., CLASS) based on the two different scoring methods?
3. What was the predictive validity of ECERS-R based on two different scoring systems?
4. What were coaches' perceived consequential validity of using ECERS-R based on two scoring systems?

Utilizing a state representative early childcare sample collected in 2013-2015, I evaluated the descriptive differences between the two scoring methods and compared the convergent validity with the CLASS tool. Moreover, I conducted a series of regressions to examine its predictive validity with child learning outcomes. Based on findings from the quantitative analysis, I gathered feedback from 13 coaches about their use of ECERS-R scoring systems to identify coaching goals and data-based decision making. Findings emerged to shed light on meta-inferences related to research, practice, and policymaking of childcare program quality assessment and improvement.

In chapter two, I synthesize the framing literature that guides my research questions. Chapter three documents the research methods and in chapter four I highlight the key findings related to each question. Chapter five provides discussion and implications based on the research findings.

Chapter Two: Literature Review

The purpose of this study is to compare the validity evidence of ECERS-R based on two scoring systems. To provide background to the reader to support the study, I extend on two features of current early childhood education programs and policies in the current landscape in the United States. These features---the QRIS and one of the primary measures used in QRIS---form the basis for this study examining the validity of the ECERS-R, with particular attention to how it is scored. This review begins with synthesizing literature around QRIS, ECERS-R, measurement theory, and ECERS-R validity.

QRIS

Over the past decade, more than 30 states have adopted child care Quality Rating and Improvement Systems (QRIS) to measure, promote, and monitor program quality (National Child Care Information and Technical Assistance Center, 2010). The political viability of establishing QRIS is predicated on the ability of a high-quality early education to set children on an upward trajectory of learning and productivity (Gordon et al., 2015). Similar to rating systems for restaurants and hotels, QRIS awards quality ratings to early care and education (ECE) programs that meet a set of defined program standards. By participating in a State's QRIS, ECE providers embark on a path of continuous quality improvement. Typically, QRISs include five components:(1) quality standards; (2) accountability measures – a process for monitoring or assigning ratings based on the quality standards;(3) connection with resources to support quality improvement;(4) financial incentives; and (5) parent/consumer education efforts (Child Care Bureau, 2007, 2010; Mitchell, 2005; National Center on Child Care Quality Improvement, 2015; Paulsell, Tout, & Maxwell, 2013; Zellman & Perlman, 2008). Below I briefly describe each element (National Center on Child Care Quality Improvement, 2013).

Standards. QRIS standards are used to assign ratings to programs that participate in QRIS, providing parents and the public with information about each program's quality. States typically use licensing standards as the starting point, or base of the system, on which higher levels of quality standards are built. Every QRIS contains two or more levels of standards beyond licensing, with incremental progressions to the highest level of quality as defined by the State. Systems vary in the number of levels and the number of standards identified in each level. The types of standards that are used to assign ratings are based on research about the characteristics of programs that produce positive child outcomes.

Accountability measures. Accountability and monitoring processes are used to determine how well programs meet QRIS standards, assign ratings, and verify ongoing compliance. Monitoring also provides a basis of accountability for programs, parents, and funders by creating benchmarks for measuring quality improvement. In most States, QRIS is monitored by the licensing agency alone or in partnership with the subsidy agency or a private entity. Most often, monitoring is conducted by separate QRIS staff within the licensing agency. Most States monitor annually, but some monitor more frequently. In particular, environment rating scale (ERS) assessments for programs participating in the QRIS are required in 20 States. Most of the States use only the ERS developed by the Frank Porter Graham Child Development Institute at the University of North Carolina at Chapel Hill. These scales are the Early Childhood Environment Rating Scale–Revised (ECERS-R), the Infant/Toddler Environment Rating Scale–Revised (ITERS), and the Family Child Care Rating Scale (FCCRS).

Program and practitioner outreach and support. Support for providers, such as training, mentoring, and technical assistance, are included in QRIS to promote participation and help programs achieve higher levels of quality. All States currently have professional

development support systems to assist coaches. These systems organize training opportunities, recognize coaches' achievements, and help ensure the quality of available training. States may use these systems to help programs meet higher professional development standards and progress toward higher QRIS ratings.

Financial incentives. QRIS use financial incentives to help child care programs improve learning environments, attain higher ratings, and sustain long-term quality. Financial support can be a powerful motivator for participation in QRIS. All statewide QRIS provide financial incentives of some kind, including higher reimbursement rates linked to the child care subsidy system, bonuses, quality grants, or merit awards; loans linked to quality ratings; and priority given to applications for practitioner wage initiatives, scholarships, or other professional development supports.

Parent/consumer education efforts. QRIS provide a framework for educating parents about the importance of quality in early care and education. Most QRIS award easily recognizable symbols, such as stars, to programs to indicate the levels of quality and inform and educate parents. Easy and widespread access to information about ratings is important. Many states post ratings on websites, while others promote QRIS through media, posters, banners, certificates, decals, pins, and other items that are displayed by rated programs.

One of the challenges related to quality measurement is validating measurement across the full range of quality and across types of early childhood settings (Zaslow, Tout, Halle, Vick, & Lavelle, 2010). Although ERS is widely adopted in many states' QRIS, its validity evidence remains inconsistent. Before systematically evaluating the validity evidence, I provided some background about ERS in the next section.

ECERS-R

The ECERS-R was not designed specifically for its current use in QRIS and other policy and evaluation efforts. For instance, it was not designed with the one and only purpose of identifying certain aspects of quality that support children' s school readiness or to be precise enough to support high-stakes decisions regarding whether programs fall above or below certain cutoffs.

Rather, the ECERS-R instrument was "based on a checklist of items for improving the quality of environments in early childhood classrooms that Harms (one of the instrument creators) had compiled during nearly 20 years of teaching and observation" (Frank Porter Graham Child Development Institute, 2003, p. 9). First published in 1980, and revised in 1998 (Harms & Clifford, 1980 ; Harms et al., 1998), the measure reflects the early childhood education field' s concept of developmentally appropriate practice, including a predominance of child-initiated activities selected from a wide array of options; a whole-child approach that integrates physical, emotional, social, and cognitive development; and highly trained teachers who facilitate development by being responsive to children' s age-related and individual needs (Bryant, Clifford, & Peisner, 1991 ; Copple & Bredekamp, 2009 ; Cryer, 1999 ; Harms et al., 1998).

Many ECERS-R items are also organized around the way in which child care center directors and teachers structure the care setting, reflecting the practitioner-focused origins of the scale. The scale developers noted that this organization makes it easy for observers to "collect information that is likely to be found under similar circumstances" (Cryer, Harms, & Riley, 2003, p. 6). ECERS-R includes six major subscales titled Space and Furnishings, Personal Care Routines, Language Reasoning, Activities, Interaction, and Program Structure. Each subscale organizes various aspects of quality within different areas of the classroom (indoor space, gross motor space, space for privacy), events of the day (meals/snacks, greeting/departing, nap/rest),

activities (blocks, music, art), and time use (schedule, free play, group time). The organization around events of the day makes it likely that items mix aspects of quality relevant for multiple domains of child development.

The scale asks observers to look for numerous features— referred to as indicators— that are attached to the scores of each item. The brief item labels do not always fully signal all of the developmentally relevant content captured by these indicators. For example, Item 10 ("Meals/snacks") contains not only indicators of nutrition and sanitation but also indicators of the amount of conversation that takes place during meals and the tone of staff– child interaction, overlapping the content signaled by the labels of other subscales such as Language-Reasoning and Interaction.

The standard scoring procedure for the ECERS-R reinforces its holistic approach and makes it difficult for researchers and policymakers to pull out specific aspects of quality and examine whether these most strongly support particular domains of development. Each item is scored on a scale with odd-value labels, from 1. inadequate quality to 3. minimal quality to 5. good quality to 7. excellent quality. In the standard stop-scoring, indicators for lower scores must be met before indicators of higher scores are evaluated. That is, observers stop scoring when they reach a response category at which an indicator is not observed. This standard scoring approach reduces response burden, in that observers do not have to consider indicators above the stopping point. It may also reflect a philosophical perspective that centers should not get credit for higher level aspects of quality that they are doing well (e.g., being warm and responsive in their interactions with children) if they are not doing lower level aspects of quality well (e.g., ensuring basic cleanliness and safety), consistent with a desire to measure the global quality of the child care environment (Clifford, Reszka, & Rossbach, 2010; Cryer et al., 2003).

The stop-scoring was not based on empirical evidence, however. The sorting of indicators into items and the placement of indicators at different scale levels was based on the scale developers' understanding of quality, based on their experiences in classrooms and understanding of the literature (Clifford et al., 2010; Cryer et al., 2003; Harms et al., 1998), rather than psychometric evidence. Specifically, research (Gordon et al., 2015) suggested that indicators placed by scale developers at lower category levels (e.g., 1 and 3) in fact reflected lower levels of an underlying dimension of quality than indicators placed at higher category levels (e.g., 5 and 7).

The stop-scoring approach challenges policymakers, coaches, and researchers who wish to isolate particular quality components in order to examine how they inter-correlate (e.g., Are some centers high in some aspects of quality, such as those that promote health and safety, but lower in others, like those that support language development?) and how they relate to child outcomes (e.g., Do aspects of quality that experts rate to be highly supportive of language development correlate more highly with language outcomes than do health-specific aspects of quality?). For policymakers, evaluators, and researchers interested in school readiness and child development, it is thus important to check whether alternative scoring approaches would produce more domain-specific measures of quality than the standard stop-scoring approach. Doing so is consistent with attempts beyond the ECERS-R to consider whether and how to develop measures of child care quality specific to domains of child development (Burchinal, Kainz, & Cai, 2011; Forry et al., 2009; Vandell & Wolfe, 2000; Zaslow et al., 2006; Zaslow et al., 2011). A focus on domain specificity can also be advantageous from a measurement perspective, as psychometricians recommend that measure development

begin by carefully defining each dimension, differentiating the dimensions from one another, and writing items specific to each dimension (Wolfe & Smith, 2007).

Meanwhile, the developers also propose an *alternative scoring* system. Specifically, the assessors are required to score all the indicators beyond the quality level score assigned to an item as each indicator can be given a score. Using the alternative scoring system can provide additional information on areas of strength beyond the quality level score and may be helpful in making plans for specific improvements and interpretation of research findings. In ECERS-3, the developers highly recommend scoring all indicators for all items when using the scale. But the stop-scoring system is still the standard rubric (Harms et al., 2015).

The use of two scoring systems (i.e., stop and alternative scoring) may have important implications in research and practices. Scoring is the foundational aspect of validity. Other advanced analyses (predictive validity, factor analysis) are based on the assumption that the scoring system is appropriate. Different scoring systems may also influence outcome interpretation and social validity. For example, preliminary research (Hofer, 2008, 2010) indicates that 25% of the childcare centers can move above one state's cutoff for more funding using alternative scoring instead of stop scoring methods. Yet, what remains unknown is whether the alternative scoring system may actually better reflect the factor structure and thus improve predictive validity of child outcomes. No research, however, systematically compares the difference of validity inferences based on these two scoring methods. It is important to investigate this question because validity claims about program quality factors based on inappropriate scoring system may not only jeopardize the quality of childcare research, but could also provide misleading information for coaches and policymakers.

In the following two sections, I will explain how scoring rules may affect other aspects of measurement validity based on Kane (2006)'s validity argument framework. Then systematically review the current literature related to ECERS-R validity to highlight the knowledge gap about scoring inference.

Measurement Theory

The quality of validity evidence is an important indicator of psychometric properties. Grounded in Cronbach's (1988) construct validity work, Messick (1989) defined validity as an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on the measurement constructs. Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on the measurement constructs (Messick, 1989). Kane (2006) further proposed an argument-based approach to validation, which involves the specification (the interpretive argument) and evaluation (the validity argument) of the proposed interpretations and uses of the scores. Kane (2006) suggests that validation tends to have two distinct but closely related usages: First. the development of the evidence to support the proposed interpretations and uses; and second. the evaluation of the extent to which the proposed interpretations and uses are plausible and appropriate. Validity inferences include for levels of inferences: scoring, generalization, extrapolation, and decision (see Figure 1).

Using this argument approach, one can exam the validity evidence at the level of *scoring*, *generalization*, *extrapolation*, and *implication* inferences. Kane's (2006) validity framework illustrates the relationships of the four levels of validity evidence and emphasizes the importance of scoring as the foundation for other advanced validity arguments. Inappropriate scoring method

may introduce systematic bias to the instrument and influence other validity evidence. Below I define each validity inference, its assumptions, and acceptable evidence related to program quality measures.

Scoring. The scoring inference connects the observed interactions with the scores given to those interactions. The scoring inference employs a scoring rule or rubric that provides guidelines for assessment. The scoring inference relies on two basic assumptions: that the scoring criteria are reasonable and that they are applied appropriately. The scoring process is free of bias, and that any statistical models (scaling, equating) employed in scoring are appropriate. Much of the evidence for the scoring inference is based on the judgment of panels of experts who develop and review the scoring criteria and procedures (Clauser, 2000). Empirical data can also be used to check on the consistency and accuracy of scoring (Clauser, Harik, & Clyman, 2000). Importantly, if the scoring rubrics reflect inappropriate criteria or fail to include some important, relevant criteria, it may systematically introduce measurement errors and affect other advanced validity arguments.

Generalization. Generalization inference refer to the extent whether the dimensions of program quality in the measure is representative of the universe score of the trait. It is important to examine evidence about how these dimensions of sampling shape scores. For observations, the goal is for score variance to be accounted for by program quality factors, for example, teacher, classroom, and so on. There will be other factors that do not shape program quality but are important to detect and quantify (e.g., observer or time of day variance). The variation that is not accounted for by program quality factors is considered error and, if large, might suggest the protocol is not working as intended. Evidence to support generalization argument is based on the reasonable effort to justify that the observation sample is representative (Kane, 2006).

Extrapolation. The extrapolation inference makes claims about the degree to which scores from the observation protocol are related to the trait. Every method of score aggregation is implicitly a scoring model and therefore, the model fit should be examined. Observation scores are aggregated to domain scores, each designed to represent a unique aspect of program quality. One way to evaluate model fit is to use factor analysis to assess the extent to which the data fit the protocol's hypothesized model structure. Evidence supporting the extrapolation inference can compare observation scores with other assessments of program quality (i.e., convergent validity). Those assessments might be measures of the same or different constructs observations measure. For example, one might hypothesize that teaching that scored high on the observation protocol should be positively related to scores designed to measure student learning. Another hypothesis could suggest that observation scores should be positively correlated with student outcomes (Kane, 2006).

Interpretation. The implication involves extensions of the interpretation to include any claims or suggestions associated with the trait. Assessment-based decision procedures typically incorporate an interpretation and a decision based on the interpretation (Kane, 2006). A semantic interpretation is evaluated in terms of its plausibility. And the evaluation of the effectiveness of a decision rule involves an application of utility theory (Cronback & Gleser, 1965). In practice, the analysis is usually qualitative. The focus in evaluation decision procedures is on value judgment (what is "desirable") and on empirical claims (about the likelihoods of various outcomes).

Measurement validation is a continuous process. The goal in designing measurement procedures is to standardize in ways that control random error effectively while introducing as little systematic error as possible (Kane, 2006). Two main threats to the plausibility of trait interpretations are trait underrepresentation and irrelevant variance (Cook & Campbell, 1979;

Messick, 1989). These two threats for a measurement tool are evaluated with the available evidence. Using Kane (2006)'s framework, I systematically review the validity evidence of ECERS-R.

ECERS-R Validity

Evaluating validity evidence of program quality measures is an ongoing process. Based on the information provided by the developers, the ECERS-R seems to be highly reliable and valid. However, empirical literature published in the recent years seem to produce inconsistent findings. I searched studies related to ECERS-R validity and coded them based on Kane (2006)'s levels of inference. Below I synthesize the validity evidence of ECERS (see Table 1).

At the *interpretation* level, most studies examining evidence for the validity of the ECERS-R have focused on its associations with child outcomes. But the results are inconsistent with many of the studies indicating weak relationships between ECERS-R and child outcomes. For example, Sylva et al. (2006) used a representative sample of 141 preschool programs in UK and results indicated that ECERS-R was a significant predictor of children's cognitive/linguistic progress. However, Weiland, Ulvestad, Sachs, and Yoshikawa (2013) found small or null associations between quality predictors and children's outcomes and that some of these relationships were curvilinear. Gordon et al. (2013) reported standardized coefficients below .10 when associating ECERS-R scores with child cognitive, social emotional, and health outcomes.

Even some significant findings are identified, the effect sizes across the studies are small. A recent meta-analysis (Burchinal et al., 2011) of similar studies suggests small effect sizes between ECERS-R and children's cognitive and social emotional outcomes, ranging from .02 to .09. Mayer and Beckh (2016) examined the validity with a German National sample and the regression analyses revealed small effect sizes for predicting child outcomes.

At the *extrapolation* level, many studies focus on understanding the factor structure of ECERS-R. Results suggest that neither a single factor nor the ECERS–R subscales are found, but rather a three-factor solution (Burchinal et al., 2009; Cassidyet al., 2005; Clifford et al., 2010; Early et al., 2006; Perlman et al.,2004; Sakai, Whitebook, Wishard, & Howes, 2003). Other reviewed literature (e.g., Cassidy et al., 2005; Clifford et al., 2005; Early et al., 2006; Frede et al., 2007; Sakai et al., 2003; Gordon, Rachel, Hofer, & Fujimoto, 2015; Gordon, Fujimoto, Kaestner, Korenman, & Abner, 2013) suggest contradicted factor structures compared to the developers (Clifford et al., 2010).

Using expect judgment, several studies focus on evaluating the face validity of ECERS-R. Contradicting with the developers' findings (Harms & Clifford, 1983), Gordon et al. (2015) surveyed 76 experts and results suggested that indicators best fit together to measure a single dimension rather than the original subscales. For example, over half of the indicators of the ECERS-R activities subscale did not fit together to define a single underlying dimension. Cassidy and colleagues (2005) found that half of the indicators on the ECERS-R measured structural quality, materials, and activities; the other half measured elements of process quality, language, and interactions. Based on the identified factors, the ECERS-R seems to capture both quality and process instead of what the developer proposed. Moreover, Sakai et al. (2003) notes that ECERS-R falls short in key components of culturally sensitive practice, such as communicating with families in their home language. They used a diverse sample in California and found that programs fell short on linguistic match could still be rated as high in quality.

At the *generalization* level, results of three studies suggested that a randomly chosen and much smaller subgroup of items provide information that is very similar to information generated by administering the full ECERS-R (Perlman, Zellman, & Le, 2004). The study of

Gordon et al. (2015) suggests that relatively few indicators capture the moderate to high range of quality.

To summarize, validity evidence at the advanced levels seem to be inconsistent. Recent research points out a fundamental problem: The mixed findings could partially be attributed to the scoring rule and the fact that within each single item different aspects of quality regarding children's basic needs are captured (Mayer & Beckh, 2016; Gordon et al., 2013; Gordon et al., 2015). Below I further elaborate on the knowledge gap at the scoring level and argue how it might affect the current validity work of ECERS-R.

Scoring Systems and ECERS-R Validity

The appropriateness of a scoring system is a fundamental validity aspect (Kane, 2006). At present, ECERS-R provides two scoring systems yet both systems are not based on psychometric evidence (Gordon et al., 2015). Using item response theory (IRT) techniques, several researchers suggest that the assumption of the stop-scoring rule may not hold (Gordon et al., 2013; Gordon et al., 2015; Mayer & Beckh, 2016). For example, Mayer and Beckh (2016) found that all items had at least one instance of disorder in the response categories according to the threshold estimates. The finding does not support the postulated ordinal structure of the categories, which assumes increasing quality from lower to higher categories. This means that empirically higher categories do not necessarily represent higher quality.

Another issue related to the current scoring system has to do with representation of variability. With the way that the ECERS-R is scored, one classroom may receive a score of a three but have many attributes that correspond to higher levels of the scale, while another classroom that also has a score of three on that same item might not have any of those higher attributes (Hofer, 2008). This fallible classroom comparison leads to another issue with the

ECERS-R 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-R scores may in actuality be very different from one another when the individual attributes present in each of those classrooms are considered (Hofer, 2010).

The third issue related to the current scoring system is temporal stability. 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). The test-retest reliability of the ECERS-R is rarely examined or reported. In fact, the only 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 (Hofer, 2010). Moreover, the length of observation, and time of day of the observation can influence the score a classroom receives and, ultimately, the funding for which such a program is eligible.

In terms of consistency and accuracy, the developers conducted a reliability study with 45 classrooms in 1997 and the result of inter-rater reliability were not satisfied. After revision and reduced the sample size to 21 classrooms, the percentage of agreement across all the 470 indicators was 86.1%, with no item having an indicator agreement level below 70%. At the item

level, the agreement was 48% for exact agreement and 71% for agreement within one point (Harms et al., 2005). In other words, the exact inter-rater agreement of the current stop-scoring system was quite low and it is quite often that two well trained raters may score at least one-point different despite observing the classroom at the same time, which may systematically introduce measurement errors at the scoring level.

The developers have attempted to develop new scoring approach (Clifford, Sideris, & Neitzel, 2012). However, the recently released third edition of the measure retains much of its existing structure and the stop-scoring rule (Harms et al., 2015). In ECERS-3, the developers highly recommend scoring all indicators for all items when using the scale, they leave the scoring decision for coaches. However, there is no psychometric evidence to compare the difference of the two scoring systems and guide the scoring decision.

To summarize, the use of two scoring systems (i.e., stop and alternative scoring) may have important implications in research and practices. Scoring is the foundational aspect of validity (Kane, 2006). Other advanced analyses (e.g., predictive validity, factor analysis) are based on the assumption that scoring is appropriate. Yet, what remains unknown is whether the alternative scoring system may actually better reflect the factor structure and thus improve predictive validity of child outcomes. The current scoring recommendation is not based on empirical evidence (Gordon et al, 2015). Although the developers mention alternate options in the instrument's instructions, they do not discuss the implications of deciding to use one method or another. If multiple ways to score an instrument are possible, it is important that those different methods offer comparable results to inform decision making. No research, however, systematically compares the difference of validity inferences based on these two scoring methods. It is important to investigate this question because validity claims about program quality factors

based on inappropriate scoring system may not only jeopardize the quality of childcare research, but could also provide misleading information for coaches and policymakers. Given the controversy and disagreement of different QRIS scoring systems across states, identifying an appropriated scoring method will have important practical implications.

Consequential Validity and ECERS-R

Consequential validity is a critical aspect at the interpretation/implication level (Kane, 2006). As shown in Table 1, however, no study directly addresses the issue of consequential validity related to the scoring systems. For classroom assessments, the goal should focus on the interpretation of performance in context rather than just paying attention to scores (Kane, 2006). The purpose of QRIS assessment is to provide valid information to support coaches' program quality decision. At the heart of the QRIS is connecting the quality improvement resources (e.g., training, coursework, coaching) with ECE programs (Tout, Epstein, Soli, & Lowe, 2015). Two recent reviews of coaching and consultation in early childhood classrooms (a key component of most QRIS) concluded that coaching is positively linked to improvements in observed quality and teacher practices (Akers & Aikens, 2011; Isner et al., 2011). Clear goals for the facets of quality and corresponding child outcomes addressed in an initiative provide an essential foundation for a quality improvement initiative (Aiken & Akers, 2011; Tout et al., 2010). Based on a review of the coaching literature, Aikens and Akers (2011) conclude that specificity of goals and clear linkages between the content of goals to the desired outcomes is an important predictor of an effective program.

Despite the fact that ECERS-R is widely used in many QRIS to support coaching goal setting and progress monitoring, no study gathers consequential validity from coaches about the assessment tools to support coaching decision making. A common approach in the field is to

conduct observations of the providers and then provide feedback to the providers with the goal of improving specific classroom practices (e.g., Cusumano, Armstrong, Cohen & Todd, 2006; De Grosso, Hallgren, Paulsell & Boller, 2010). One way is to use a formal, observational tool such as the ECERS-R to assess the routines, materials, and basic foundational level of quality in a program (e.g., Wesley et al., 2010; Palsha & Wesley, 1998; Boller et al., 2010). Findings indicate that an environmental assessment tool can be used at the beginning of an intervention to assess program needs and to inform goal development by conducting a joint assessment with the consultant and provider (e.g., Wesley et al., 2010). However, no study was found that attempts to understand the benefit and affordance of using standardized assessment to support coaching observation and goal setting.

The use of different scoring systems may have important implications for coaching and quality improvement. Emerging research suggests the importance of promoting a culture of positive change and assessment of strengths and needs (Tout et al., 2015). A core feature of QRIS is using data to make informed decisions about quality improvement priorities, and on developing the capacity for ongoing assessment and improvements (Mitchell, 2012; Wiggins & Mathias, 2013). While standardized assessment has growing appeal to policy makers and researchers, we do not know whether the current assessment tools like ECERS-R are fulfilling coaches' needs of assessing program and evaluating progress. In particular, we do not know which score system might better support the needs of promoting a culture of positive change and assessment of strengths and needs (Tout et al., 2015). Gathering consequential validity data and practitioner wisdom from coaches could help inform assessment development.

Given the important role the ECERS–R has played in both research and policy in the United States and around the world, it is critical to understand how two scoring systems may

affect the validity inferences at scoring and decision making levels. In the next chapter, I report the research methods that aim at understanding the validity differences based on the two scoring systems, and how coaches' perceive consequential validity of using ECERS-R to make decision based on two scoring systems.

To fill this high-stake knowledge gap and inform childcare quality improvement research, practices, and policy making, I conceptualize a mixed method study to examine the validity of ECERS-R. The purpose of this study is to compare the validity evidence of ECERS-R based on two scoring systems. The research questions include:

1. What were the score differences between the two scoring methods?
2. What was the convergent validity between ECERS-R and another quality measure (i.e., CLASS) based on the two different scoring methods?
3. What was the predictive validity of ECERS-R based on two different scoring systems?
4. What were coaches' perceived consequential validity of using ECERS-R based on two scoring systems?

Chapter Three: Methods

Evaluating an instrument' validity evidence is an iterative process. Given different evidence needed to support validity arguments at different levels (Morse, 1991), I decided to use sequential mixed method to maximize validity evidence confidence (Jick, 1979). Specifically, I used quantitative analysis to examine inferences at the scoring and extrapolation levels. Then I employed qualitative interviews to gather consequential validity data. Findings from the quantitative analysis helped inform potential questions needed to ask from the coaches. Likewise, findings from the qualitative component may reveal the limitations of ECERS-R and meta-influence can be drawn to inform policy making, practitioner decision, future validity research. Next I layout the detailed information for the quantitative and qualitative components.

Quantitative Component

Data source. Through a data-sharing agreement I obtained a secondary data set in which raters had scored the ECERS-R with both the stop-scoring and alternative-scoring rules. The data set was part of a state early childhood program quality assessment and improvement project conducted at University of Washington.

Setting and participants. Early childhood programs (N=120) in a U.S. Northwestern state were included in the study. The program types included federal funded (i.e., Head Start), state funded programs (i.e., ECEAP), and private family child care centers. The sample included both infant/toddler and preschool classrooms across all regions of the state. The sample achieved adequate representation across Washington State, with sites in each of the seven Child Care Aware regions. There were 16% sites in Central Washington, 15% sites in Eastern Washington, 15% sites in Northwest Washington, 22% sites in King County, 9% sites in Pierce County, 10% sites in Southwest Washington, and 13% sites in the Olympic region.

Children in each site were randomly selected for participation—up to eight children (four boys and four girls when possible) per site. The mean age of the children (*N*=531) was 4.3 (*SD*=0.6) years old, 51.2% of them were male, and 22.6% of the children did not speak English as their primary language. The children in the study sample were from various racial backgrounds (Table 2), with the largest group being white (48.6%). Additionally, 17.3% of the children were of Latino ethnicity. Teacher/provider reports indicated that 12.1% (64) children had been referred for special education, and 10 parents reported that their children had developmental delays. The children these programs served reflected a wide range of socio-economic status. As shown in Table 2, about 36% (*n*=190) of the children's families received subsidy. About 49% of the parents had a bachelor's degree or above while the rest had less than 9th grade to an associate's degree. Approximately 48% of children's families had $50,000 annual income or less (Table 3).

Instruments. The battery of measures used for child and program assessments included a variety of observation and standardized assessments. Child assessment data were collected across learning domains (see below) in fall 2014 and again in spring 2015 (about 150 days later) to determine children's learning over time. Below I describe the instruments (also see Table 4) included for this study:

Program quality. The Early Childhood Environment Rating Scale—Revised (ECERS-R) (Harms, Clifford, & Cryer, 1998) assessed quality of preschool classrooms on the following subscales: Space and Furnishings, Personal Care Routines, Language Reasoning, Activities, Interaction, and Program Structure. The total scale consisted of 37 items and was used for observations in classrooms with children from 30 months to five years of age.

In addition, the Classroom Assessment Scoring System Pre-K version (Pianta, La Paro, & Hamre, 2008) was administered for the same sites as part of the state QRIS system. The CLASS is a widely used process quality tool that assesses classroom practices by measuring teacher-child interactions and material use. It is available in multiple versions and the tool has been linked to student achievement and development and has been validated in more than 2,000 classrooms. The CLASS Pre-K version is appropriate for classrooms serving children 3-5 years of age and is comprised of three domains: Emotional Support, Classroom Organization, and Instructional Support. The CLASS measure is highly convergent with teacher self-efficacy (Pakarinen et al., 2010). Findings related to predictive validity are mixed. La Paro, Wiliamson, and Hatfield (2014) suggested that children in classrooms with higher levels on the CLASS-Toddler domains of Emotional and Behavioral Support as well as Engaged Support for Learning were reported to have fewer behavior problems. However, results of another study (Bell et al., 2012) suggested the relationship between CLASS secondary measure and student outcomes were tenuous.

Early reading. For preschool-age children, letter word knowledge was measured using the Woodcock–Johnson III Tests of Achievement (WJ III; Woodcock, McGrew, & Mather, 2001) Letter-Word Identification subtest, which assesses a child's ability to identify letters and words. For Spanish speaking children, the Batería III Woodcock-Muñoz was used. The WJ-Letter Word Identification are standardized instruments with a mean of 100 and a standard deviation of 15. The median reliability coefficient alphas for all age groups for the standard battery of the WJ III ACH for tests 1 through 12 ranged from .81 to .94. The internal correlations of the entire battery are consistent with relations between areas of achievement and between areas of achievement and ability clusters. Growth curves of cluster scores in the technical manual illustrate expected

developmental progressions, with steep growth from age 5 to 25, with a decline thereafter (Woodcock, McGrew, & Mather, 2001)

Early science. The Lens on Science (LENS; Greenfield et al., 2009) is an adaptive computer-based instrument that assesses preschool children's content and processing skill in science. The LENS yields scores between -3 and 3 with a mean of zero and a standard deviation of one. Reliability and validity data collected with the SDA are reported demonstrating both high person reliability (.93) and item reliability (.98), predictable correlations with related measures, growth in science ability across the preschool school year and sensitivity to detect the positive impact of a classroom based preschool science intervention.

Early writing. The Early Writing Assessment (EWA; Puranik, 2011; Puranik, 2012) measures early writing development and asks preschool children to write their names (EWA Name) as well as two consonant-vowel-consonant words (EWA Word) from dictation. The early writing rubric for the EWA Name task is scored on a scale of 0-8, while the range of possible scores on the EWA Word is 0-18. No study reported validity information about this measure.

Executive function. Preschool-age children's executive functioning was assessed using the Head Toes Knees and Shoulders (HTKS; Ponitz et al., 2008), a measure of behavior regulation, and effortful control specifically. Effortful control is the ability to stop doing something (inhibit a response) and do something else instead. The HTKS is a three-part activity with 60 points possible. Previous study indicated inter-rater reliability for total score and self-correct responses was 66% and 75% respectively (Ponitz, McClelland, Matthews, & Morrison, 2009).

Early math. Preschool children's early math knowledge and skills was assessed using the Tools for Early Assessment in Math (TEAM; Sarama, Clements, & Wolfe, 2010). Children participated in a short form version of the TEAM, which included 20 questions and a stop rule.

Using Rasch-model analysis, the developers suggested the construct validity of these developmental progressions was supported.

Receptive language. The Peabody Picture Vocabulary Test, Fourth edition (PPVT-4; Dunn & Dunn, 2007) is an individually administered instrument measuring the receptive vocabulary of preschool children. The PPVT–4 measures understanding of the spoken word and thus assesses receptive vocabulary levels. For Spanish speaking children, the Test de Vocabulario en Imagenes Peabody (TVIP) was used. The PPVT-4 and WJ-Letter Word Identification are standardized instruments with a mean of 100 and a standard deviation of 15. Reported test-retest reliability is .93. Convergent validity with the PPVT-3 is .84. The trend of average performance across age was compared with the profile of growth and decline in crystallized ability reported in the research literature (Dunn & Dunn, 2007).

Social-emotional. The Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2000). yields Internalizing, Externalizing, and Total Problems scales as reported by parents and teachers/providers. The CBCL has a mean of 50 and standard deviation of 10. The test-retest reliability is high at .85. For interrater agreement, the mean was .61. The content validity of the scale is supported by findings that nearly all items discriminated between referred and nonreferred children by the extensive process. The construct validity of the scale is supported by concurrent and predictive association with a variety of other measures (Achenbach & Rescorla, 2000).

Data collection. Data was collected in three phases (October 2014-February 2015; February-May 2015; and March-July 2015). Assessors were extensively trained to observe classrooms and provide ratings on the ECERS-R and CLASS. Data collectors were held to a

stringent threshold of 80% reliability for the CLASS and engagement measure and 85% for the ECERS-R to establish initial reliability before conducting classroom observations. The data collection team was led by six anchors who established reliability with the instrument authors during three live classroom visits in summer 2015. Average reliability percentages on these author/anchor visits were 92.7%. Data collectors, who were scheduled internally by research team staff, arrived at sites "unannounced." Program directors were given a range of dates spanning 2-3 months, during which they could expect the classroom visits to take place, but they were not provided any details as to the specific date of the observation. Data collection occurred over the course of two calendar days with the first day focused on ECERS-R and the second day on CLASS. Upon arrival, the data collector checked in with an on-site contact who then led them to the appropriate classroom. The length of the observation depended on the measure being used. Prior to the unannounced visit, data collectors were briefed to refrain from interacting with children or teachers to avoid any unnecessary interruptions to instruction.

For the CLASS, the team conducted four 20-minute observation cycles, as recommended by the instrument authors. These assessors coded and scored the observational measures after each site visit before submitting documents to office staff, who then conducted a quality check on the data. ERS data were collected during visits by reliable external assessors. Assessors were extensively trained to observe classrooms and provide ratings on the ERS measures. The ERS observations lasted for approximately three hours, during which time the data collectors limited their interactions with children and providers. At the conclusion of each site visit, protocols were submitted to office staff who then conducted a quality check on the data.

The data collectors were trained to reliably conduct individual child assessments throughout fall 2014 and again in spring 2015 to determine children's learning and development

over time. Preschool-age children (3 years and older) participated in assessments measuring language, letter word knowledge, early writing, early math, early science, and executive functioning. Social emotional skills were assessed via teacher report on a standardized measure (i.e., CBCL). Staff members, who visited sites in teams of two to three, were scheduled based on assessment needs per program type and composition (FCC/CCC, I-T/PK, English/Spanish). Teachers and providers guided staff to a well-lit and distraction-free area to engage in activities with randomly selected children. The study team administered assessments to children individually, paying particular attention to level of engagement and providing breaks as needed. Before leaving the site, the assessors offered all teachers and providers their choice of children's books to keep as a token of appreciation. More information about data collection is available in the project report (Soderberg, Joseph, Stull, & Hassairi, 2016).

The subscale Parents and Staff (items 38–43) in the ECERS-R was excluded from the analyses to secure comparability of the results with the publications (e.g., Gordon et al.,2013; Mayer & Beckh, 2016). The removal of the Parents and Staff subscale is also consistent with prior research (e.g., Clifford et al., 2005; Mashburn et al., 2008) and the new version ECERS–3 (Harms et al.,2015).

Data analysis. I first checked the accuracy of the data set and examined the descriptive statistics to identify potential outliers. To handle the missing data in the demographic variable (i.e., subsidy), I treated it as missing at random and chose multiple imputation to stimulate the missing data. All the other missing data in the child outcome variables were list-wise deleted. Next, to make the two scoring systems comparable, I converted the stop scores (range from 1 to 7) into ratio scores (0-100%) through the formula of [(n-1)/6]*100%. For example, a score of "1" in the stop score scale would be (1-1)/6=0, and "4" would be [(4-1)/6]*100%=50%. The

alternative score was calculated by adding all the obtained scores and dividing by the total number of indicators and multiplying by 100%. For example, if a site got 7 items out of the 14 possible indicators, the alternative score would be 50% (7/14 *100%).

For research question 1, I conducted descriptive analysis to compare the differences between the two scoring systems at the subscale and total score levels. I also conducted paired sample t-test to statistically compare the mean differences of the total and subscale scores based on the two scoring systems.

For question 2, I used zero-order correlation to examine the relationships between CLASS and the ECERS-R based on the scoring systems. I compared the results at the whole scale and subscales levels.

For question 3, I used two level hierarchical linear modeling (HLM) with children representing level-one and program representing level-two units. A two-level model was specified because this study is concerned with quality at the program level and corresponding child performance with the environment. As compared with traditional unilevel methods (e.g., analysis of variance and multiple regression), the more complex analysis method accounts for dependencies among student scores due to school membership, allowing for valid inferences to be drawn about relationships between student outcomes and school-level predictors without violating the assumption of independence (Raudenbush & Bryk, 2002). HLM provides the best framework for examining the relationships between multiple predictors of quality and child outcomes on individually administered posttest measures, after controlling for pretest ability and child characteristics. I created z scores for all the variables to make it easier to interpret the results.

First, an examination of the unconditional models for each of the outcomes was conducted to

partition the variance into two sources: within programs and between programs (see Model 1). Next, I included the pretest, individual ECERS-R subscale, and three child level variables (gender, primarily language, and subsidy to provide more accurate estimates (see Model 2). These parameters, $\beta 00\ to\ \beta 05$, vary across programs in the levele-2 model as a function of a grand mean and a random error (model 3-8).

$$Yij(outcome) = \beta 00 + eij \quad \text{Model 1}$$

$$Yij(outcome) = \beta 00 + \beta 1j\ (ERSsubscale) + \beta 2j\ (pretest) + \beta 3j\ (male) + \beta 4j\ (English) + \beta 5j\ (subsidy) + eij \quad \text{Model 2}$$

$$\beta 00 = roo + uoj \quad \text{Model 3}$$

$$\beta 1j = r10 + u1j \quad \text{Model 4}$$

$$\beta 2j = r20 + u2j \quad \text{Model 5}$$

$$\beta 3j = r30 + u3j \quad \text{Model 6}$$

$$\beta 4j = r40 + u4j \quad \text{Model 7}$$

$$\beta 5j = r50 + u4j \quad \text{Model 8}$$

I used the Statistical Package for the Social Sciences (SPSS; SPSS Inc., 2006; 1989–2004) to compute descriptive statistics. In all multilevel analyses, I used the nlme package in the R programming software (R Development Core Team, 2008).

Effect size estimates were calculated as the model-estimated treatment slope coefficient divided by the approximate standard deviation; the approximate standard deviation is computed by multiplying the model-implied standard error with the square root of the number of schools in the study. The simple pooled standard deviation from the observed data was not used because it does not account for the lack of independence of students' scores within schools, nor does it account for missing data. I termed these predicted effect size estimates as d for convenience—

similar in interpretation to Cohen's d. According to Cohen's (1992) recommendations, the magnitude of the associations was evaluated as being small with standardized coefficients of about .10 or less, medium with standardized coefficients of about .30, and large with standardized coefficients of about .50.

Qualitative Component

Research design. How coaches utilize the information from the measure and make decisions is an important aspect of consequential validity (Kane, 2006). To gain coaches' perspectives, I used a qualitative interview methodology (Merriam, 2009) to investigate the perspectives—that is, the values, beliefs, and attitudes—held by coaches with regard to their views about using ECERS-R and related assessment tools to support data-based decision making. Interview is a helpful data collection tool when one cannot observe behavior, feelings or perceptions (Merriam, 2009). Also I elected this methodological approach because of the hypothesis-generating, rather than hypothesis-testing, purposes of the study (Glaser & Strauss, 1967).

Participant recruitment. I used criterion-based selection to ensure the participants aligned with essential features and adequately addressed the research questions (Merriam, 2009). The criteria of coach participants were as follow: a. providing coaching to early childhood programs for at least two years; b. having experience in using ECERS-R and/or other program quality measures in Washington state. Based on the inclusion criteria and to maximize the variation (Merriam, 2009), I selected a list of potential participants for a focus group, including the current coaching staff at the Childcare Quality and Early Learning Center for Research and Professional Development (CQEL), and also the coaching professional workforce email list available at the center. The IRB approval was obtained before data collection. A total of 13 state

certified coaches from across the state participated in my study. Table 5 reports the working experience, and other demographic information.

Setting. The interviews were conducted at the coaches' office or a place at their convenience. Phone call or distance interviews were also used if the interviewee preferred. The interview lasted about 45 minutes to 1 hour.

Instrument. The researcher is the primary instrument for qualitative inquiry (Merriam, 2009). All interviews were guided by a semi-structured protocol (see Appendix A) with questions about experience using the assessment and decision making, such as Tell me about your experience using the ECERS-R ." Interviews were tape-recorded and transcribed. Transcripts were entered into a web-based platform for qualitative and mixed-method data analysis (http://www.dedoose.com).

Data analysis. In qualitative research, data analysis occurs simultaneously with data collection (Merriam, 2009), providing opportunities to develop new lines of questioning, member-check themes, and refine the conceptual framework. I used the standard procedures for inductive data analysis (Charmaz, 2002; Strauss and Corbin, 1997). I began my analysis by reading through the interview transcripts and identifying text segments of potential relevance to my research questions. Each of these segments was tagged using low inference descriptors such as "scoring interpretation" or "item feasibility." Then I identified and defined codes (see Appendix B) emerging from individual analysis to be formalized and used in light with the theoretical framework. The remainder of the interviews were then coded, with some codes combined, others modified or deleted based on their perceived value relative to my research questions. A set of interpretive categories were developed through this process which were used to aggregate coded data segments and which became the basis for further analysis. These

categories were then used as a basis for developing a series of data displays (Miles & Huberman, 1994) organized by a. item feasibility and cultural appropriateness, b. scoring preference and reasons, c. score interpretation, and d. decision making. I then met with the research committee members to discuss the analyses and integrated the categorical data into larger and more interpretive summaries. These summaries were used to develop three cross-case themes (Ryan & Bernard, 2003) described in the qualitative findings below.

Chapter Four: Findings

Quantitative Component

I conducted a series of analysis aligned to relevant research questions. Findings are presented by each of the four research questions.

Question 1. What were the score differences between the two scoring methods? As shown in Table 7, the mean difference between the alternative and stop scoring systems was statistically significant $p<.001$. The mean ECERS-R score for the stop score was only 46.31 (*SD*=11.36, ranging from 23.41 to 73.82), while the mean ERS alternative score was almost 33 points higher at 79.16 (*SD*=7.69, ranging from 56.59 to 90.77); $t(119)=50.79$, $p<.001$.

This pattern was observed at the subscale level. Specifically, the mean stop score for Space and Furnishings was 40.52 (*SD*=10.43, ranging from 20.83 to 77.08), while the alternative score was 76.79 (*SD*=7.50, ranging from 57.17 to 94.44); $t(119)=44.46$, $p<.001$. The mean stop score for Personal Care was only 15.42 (*SD*=8.24, ranging from 2.78 to 41.67), while the alternative score was 74.83 (*SD*=8.63, ranging from 50.24 to 90.28); $t(119)=72.57$, $p<.001$.. The mean stop score for Language Reasoning was 52.13 (*SD*=16.28, ranging from 16.67 to 91.67), while the alternative score was 76.96 (*SD*=10.29, ranging from 47.41 to 95.46); $t(119)=25.93$, $p<.001$. The mean stop score for Activities was 52.52 (*SD*=14.16, ranging from 26.67 to 94.44), while the alternative score was 77.91 (*SD*=9.17, ranging from 55.67 to 97.22); $t(119)=26.63$, $p<.001$. The mean stop score for Interaction was 56.35 (*SD*=23.30, ranging from 6.67 to 100.00), while the alternative score was 74.22 (*SD*=18.90, ranging from 10.00 to 100); $t(119)=16.92$, $p<.001$. Lastly, the mean stop score for Program Structure was 60.94 (*SD*=19.63, ranging from 5.56 to 100.00), while the alternative score was 85.89 (*SD*=11.42, ranging from 45.15 to 100.00); $t(119)=18.91$, $p<.001$.

Question 2. What was the convergent validity evidence between ECERS-R and CLASS based on the two different scoring methods? I first examined the zero-order correlation between the two scoring systems (see Table 6). Results suggested that the correlation of mean ECERS-R scores between the two methods was quite high at r=.90, p<.01. And the correlations at the subscale levels were also very high, with Space and Furnishing (r=.752, p<.01), Personal Care (r=.665, p<.01), Language Reasoning (r=.899, p<.01), Activities (r=.858, p<.01), Instruction (r=.882, p<.01), and Program Structure (r=.865, p<.01).

As shown at Table 8, the mean correlation between ECERS-R and overall CLASS was quite similar using the two scoring systems. Specifically, the correlation based on alternative scoring was r=.537, p<.01, while the correlation based on stop scoring was r=.539, p<.01. At the subscale level, the correlation pattern appeared to be similar as well. Specially, the correlation between mean ECERS-R and CLASS Emotional Support subscale was at r=.537, p<.01 using the alternative scoring system, while it was r=.529, p<.01 with the stop scoring system. The correlation between mean ECERS-R and CLASS Classroom Organization subscale was at r=.539, p<.01 using the alternative scoring system, while it was r=.558, p<.01 with the stop scoring system. The correlation between mean ECERS-R and CLASS Instructional Support subscale was at r=.310, p<.01 using the alternative scoring system, while it was r=.303, p<.01 with the stop scoring system.

Question 3. What was the predictive validity of ECERS-R based on two different scoring systems? Table 10 shows observed baseline and end-of implementation descriptive statistics for outcomes. For descriptive purposes, inter-correlations among variables included in statistical analyses are provided in Table 11. Table 12 provides the outcome summary of the fixed effect (individual outcomes were presented at Table 13-21). No statistically significant

main effects were identified on any child outcome measures for overall ERS scores, but significant main effects were identified on three child outcome measures for specific subscales of the ERS. For PPVT outcome, statistically significant main effect for space and furnishing was evident (*Coeff*=0.11, *SE*=0.04, *p*=.013, *d**=0.30) for the alterative scoring, while no statistically significant main effect was observed for the stop scoring. For TEAM outcome, statistically significant main effect for space and furnishing was observed (*Coeff*=0.07, *SE*=0.03, *p*=.039, *d**=0.25) for the stop scoring, while no statistically significant main effect was observed for the alternative scoring. For EWA-word outcome, statistically significant main effect for program structure was observed (*Coeff*=0.09, *SE*=0.05, *p*=.042, *d**=0.24) for the stop scoring, while no statistically significant main effect was observed for the alternative scoring. For all other outcome measures, no statistically significant main effect was observed for either the stop or alternative scoring systems.

Qualitative Component

The purpose of the qualitative component was to understand how coaches perceived the ECERS-R and made data-based decision with program providers. Three themes emerged based on the 13 coach interviews and guided by the theoretical framework at item, scoring, and decision making levels: (1) program providers felt frustrated to comply with some quality indicators due to item feasibility and adaptability; (2) coaches preferred the alternative scoring system as it provided more information, helped identify specific goals, and cultivated a strength-based coaching partnership; (3) data-based decision making was grounded in valid information from the tool, coaches' knowledge about the tools and the programs, and the ability to interpret and negotiate with program providers. Table 23 shows how the themes are supported by each

participant. Detailed information about each theme is delineated below. Please note that I used "…" in quotes to omit text that are redundant or self-repeated.

Theme 1: Program providers felt frustrated to comply with some quality indicators due to item feasibility and adaptability. Before discussing the scoring systems, participants first shared their thoughts about the assessment items. Many coaches acknowledged the usefulness of the tool. As coach Carta mentioned: "the ECERS-R assessment was my bible and it did help guide her classroom arrangement and bring up the program quality." However, coaches suggested that some quality indicators were hard to comply with due to two main reasons: the items were not feasible to implement due to logistic reasons such as time constraints; the items could not be adapted to local cultures and linguistic context. The subthemes were further elaborated as follow.

Item feasibility. Coaches suggested that many program providers didn't think it was feasible to implement some of the quality practices, such as hand washing, table sterilizing, and diapering. Providers felt defeated and they thought the items were too picky and not feasible to implement. They would get low scores (1=inadequate) no matter how hard they tried to meet the requirements. As one of the coaches shared.

> I know across the board a lot of people scored low in the health and safety area because some of the standards are very strange and it is frustrating. I mean I hear providers say how frustrated that no matter how hard they try, even though their licenser say they are doing everything fine and then they get a 1. You know, we can't win (Tara).

Participants further explained that they thought some practices were not feasible due to several logistic reasons, such as time constraints, physical barriers, program policy and philosophy. As a coach pointed out the time barrier:

> There are things like hand washing practices that would take up a substantial portion of the day when you are trying to use a curriculum with fidelity. That is really difficult and there is a lot of pressure for the teachers (Carta).

In addition, another coach pointed out that some programs could not meet the health and safety requirements because of physical barriers:

> We have a classroom with 18 kids and only one sink …The teacher has to rush through. If those kids have to stand in line for every single kid to wash their hands for 20 seconds, then they will be standing in line for more than 40 minutes…It is almost impossible to be able to meet the ECERS requirement to get a high score and that is very common. Some classrooms are lucky to have three sinks but that is because they have two bathrooms and each has one sink. But then we are still breaking the rule because those sinks do not get sterilized, and the kids go to the bathroom and they wash their hands and sit down to eat. So they are still going to fail (Rayann).

Furthermore, providers suggested that program policy was another barrier. For example, Rayann shared a case where the program policy prevented her from meeting the assessment requirement:

> Some of the scoring we do not have control for because we have partnership sites with the school district. Things like the playground and the play materials, which we have no control over that. And so that is going to automatically bring our score down. And then also we have employees like para-educators and teachers that are in our partnership site too. So the way they treat kids is way different than the way we do. So our score goes down because they tend to be a little harsher in the way they are talking and engaging with the kids. They may force a kid to the circle when we won't. So those kinds of things score down too. And they do not have to follow the health and safety requirement like we

> try to do so I have been in a classroom where they just do not wash the table and they do not wash hands and they do not understand why that is a big deal (Rayann).

Participants suggested that some providers were skeptical about the quality practices due to conflict with program philosophy. For example, several coaches mentioned that dramatic play and using block were not applicable in Montessori programs. Anne shared her experience of using ECERS-R to work with two Montessori programs:

> The tool doesn't blend with Montessori at all. I worked with two fabulous Montessori sites and they only rated a 3 because the ECERS-R tool. They want them to have dress-up and blocks and all that. I understand the importance of blocks, but why specifically animal and people? Show me the research that those things are more important. I could be wrong but I do not believe there are a lot of research about that (Anne).

Amy was a Montessori teacher before taking the coaching position and she had similar challenges that Anne encountered. She felt frustrated to have Montessori programs be labeled as "grandma daycare" and restricting other forms of practices:

> Practical life has a lot of dramatic play experience or role play experience and doesn't necessary have to have the clothing but the children are still getting the social emotional needs met. It is just in a different way than this ECERS-R tool allows for. It is in this way restricting Montessori sites or Reggio Emilia program or Waldorf for example... When you know, the philosophy is proven to be effective. It has been research-validated. Could they waive certain requirement, not in health and safety, but just the activity area? Maybe it would help them recognize they are a high quality program. But there is previous bias, you know, grandma daycare, and that frustrates them. Because there is a big difference

> between them and "grandma daycare." But they still get a low 3 and it is only because this tool doesn't take it into consideration in terms of different program philosophy (Amy).

***Cultural and linguistic adaptation*.** While it was helpful to clearly define and quantify the items, such description made coaches felt like it was too restricted and this was the only way to do things without considering the cultural and linguistic adaptation. Some participants suggested that the ECERS-R tool focused too much on materials rather than actual culturally responsive practices. And they argued that the assessment tools should have more specific guidance on cultural and linguistic adaptation. As Tara pointed out the needs for culturally responsive practices:

> I haven't observed where they actually honor someone's culture or even acknowledge their culture specifically. You know some of the centers I go in have stuff on every culture which is inclusive, but it almost doesn't honor the individual children culture. I mean they try to address it by adding what they think is a food from their culture, but I do not think it is done intentionally. I mean for Mexican child they put taco on and they think that is their culture (Tara).

Anne worked with programs that served different cultural and linguistic backgrounds and she echoed the importance of cultural adaptation that mirrors diverse practices.

> In general, it is not a very culturally responsive tool. If you work with the family child care providers and they are from Samoa. They eat on the floor with their hands. They do not wash their hands every time after they take a bite. Or the child displays that...I have never been at a Muslim home that has a bunch of stuff on the wall. That is something I really like to see where the research. And some of the cultural books would be crazy on

> what the world is actually going on. Some of the supervision thing I do not think make any sense when you are trying to instill independent in children but then you can't see everybody in the bathroom all year. So you know especially in the family child care centers or if you are in the building where the bathroom is down the hall. I just do not think, we want the kids to be independent, but we actually stop people from having the kids to be independent (Anne).

Another participant suggested the importance of providing more room for adaptation when considering the family childcare centers context:

> One thing that is important is to have more adaptation. For example, the family child care centers, the toys are going to be very different because of financing, supervision, that are very different than a child care center offered to the children. And sometimes just understanding the backgrounds of the different cultures that are applying this tool or trying to understand this tool (Nancy).

Family child care providers felt frustrated to implement the practices with the intension of pushing them to look like a center instead of appreciating their unique family cultures.

> The family child care providers feel very strongly that the push is to make them be more like the center and I feel that from every single one of them. Because there are components in family child care that provide reasons for people to go to child care programs instead of centers because they want it to be family-like. That is what they are looking for. And when you are using a tool like FCERS, it makes it very much like have to be a center. And the providers do not like that (Anne).

Culture also influenced people's language and communication. Participants highlighted the importance of knowing the appropriate communication practices in context like greeting parents, social interaction, and table manners before scoring:

> Culture is a huge issue…In some cultures, parents do not greet teachers when they are busy working because that is considered rude. So when they drop off their child, saying something with the teachers can be very challenging because you are just there and do not want to interrupt teachers… And that is just a sign of respect. But I know some of the sites are getting low scores in welcoming and departing but it may be just because of the culture piece. Knowing the practices and culture that are comfortable for them is critical before scoring. Another thing that is always been challenging is the language and conversation for teachers and families. For example, a conversation is expected when you are eating. But there are some groups that is the norm. And you sit and eat and finish your food and then go to the next activity. And I know a few sites get a low score in language and conversation because of it. Knowing what space you are going into is huge important (Nancy).

Assessors needed to recognize different environments and understand the intention behind the language when evaluating the relationship rather than expecting only one form of communication:

> I have been in where the conversation is still enduring and there is a minimal relationship between the teacher and the student. It can come off as more straightforward and short. But the child recognizes there is love behind the word too. When the teacher will be rated very low in the tool despite the intention. Like the marker in the tool asking whether there is a positive relationship with the kids which there is. But because of vocabulary and

> language is different than what is expected, the way it would look like in the traditional classrooms, it is not seeing as still being a loving relationship (Nancy).

Moreover, participants pointed out the importance of adapting guidelines on book choice that are culturally responsive:

> In Native American culture, they hunt, they fish, you know that is what they have done in the past. And that is in their culture and history. And they teach their children in their books, even some difficult stories. Some centers are faith-based. They read Bible stories and that can be violence (Janet).

To summarize, participants suggested that program providers found it difficult to comply with some quality indicators. Some quality indicators lacked feasibility, cultural adaptation, and compatibility with program philosophy and facility. Program providers felt frustrated to follow the quality practices despite putting in a lot of effort.

Theme 2: Coaches preferred the alternative scoring system as it provided more information and cultivated a strength-based coaching partnership. Coaches preferred to use the ECERS-R alternative scoring system to help interpret the program strengths and areas of improvement. Three subthemes emerged: 1. The alternative scoring system gave coaches more information beyond just numbers; 2. The specific information helped coaches better identify goals, monitor progress, and provide specific feedback; 3. The alternative score system helped coaches cultivate a strength-based coaching partnership.

1. The alternative scoring system gave coaches more information beyond just numbers. While the numeric score was important to quantify and determine program quality, coaches advocated that we all should look beyond numbers. They preferred specific notes and information from the assessment tools instead of just numbers.

> I look all the way up. So if a site scored at 2 but sometimes they may not have something in the 3 category but at the end of the day they have something in the 7. Ok they are 2 but I still want to see what else they have. Because then I know if they could just add a few more things, they could have reached the 6. So for me, I use the alternative score and go a little bit further just to know if they actually have other components too. This helps me to have more information as a coach so then I know ok now I know how to support them. (Yan).

In addition to the alternative scoring system, coaches preferred detailed notes from the data collectors to better understand the program and practices. And they learned that pure numbers would not help communicate well with program providers. Merely having the numbers was like a "guessing game." Nancy talked about the challenges of relying only on numbers:

> Both the CLASS and ECERS-R scoring reports are pretty challenging. The CLASS gives you more information about interaction because it is just like what kind of open-ended questions the teachers asked. So it feels more subjective. And the ECERS-R seems more objective like you are supposed to have these many art materials. I think being clear would help feedback. Like this one report I am looking at, they have a 6 in the fine motor but they have a 4 in music and a 2 in the block. I have no idea how they got a 2 in the block. You know they are able to get a 6 area, which means they are able to provide adequate materials. They are most likely to provide materials for other areas. So being able to say because they are not having the access in the block area, or because they did not have enough height in the block, or because there isn't enough area for the children to use the block. Now you know you have 6 in one area and a 2 in another area. It does not give you any information on how to improve (Nancy).

2. The specific information helped coaches better identify goals, monitor progress, and provide specific feedback. All coaches in my study suggested that the alternative scoring system helped them gain a more comprehensive picture about the program strengths and areas of improvement. Using the alternative system helped coaches better identify goals, monitor progress, and provide specific feedback. The coaches strongly wished that the data collectors would provide more detailed notes at the item level and not having providers fixated with numbers.

> It used to say that we have, we only get the number and there is not much information, and that is not helpful at all. Often time the provider want to piece together to see what action that they did mark them down. And I think if we were trying to improve the practices, giving the specific detail feedback instead of just the score would help. Because right now it is like a guessing game about what they did that make them get the low score. So having more information for sure about what tone they heard, which materials were not acceptable, and what materials were acceptable. Giving more concrete details about how specific they can improve would help (Nancy).

Participants noticed that data collectors tended to not provide notes on items that scored 4 or above. But they insisted that notes on the higher level were equally important if we wanted to promote continuous program improvement.

> I know it is a little bit time consuming but 4 and above…really I have noticed that teachers said I wished there is more clarification on why we score a 4 and what we mark off on the 5…For sites what we have been working with for a long time, I may guess what they might have been marked off on in that column on the 5. But it would be nice to have the feedback on the 4. That is something I and another coach noticed when providing feedback (Janet).

Using the alternative score system also provided coaches with more objective information:

> Oftentimes providers tend to think that they are at a certain level and they are not. In our region, we try to be transparent as far as where we think they are at and where they could be their potential. I think having a training on teaching the coaches to establish challenging conversation would be great. I think giving more detail in the rating report is definitely be a huge support because then they know what they are speaking of. They can call out specific examples of that day (Daisy).

3. The alternative score system helped coaches cultivate a strength-based coaching partnership.

The alternative scoring system cultivated a strength-based coaching partnership. The alternative scoring system might be just a simple choice or preference. But from coaches' perspectives, it represented the idea of recognizing and reinforcing providers' good practices.

> I score all the way up and that way is not so daunting to them. Let's say there is just one little thing that causes them not meet the 3 criteria. Then I can say ok if you just fix this one little thing then look at where you will be. That helps a lot, especially in a coaching role. It is more encouraging for the teacher and they say oh yeah I can do that and they can create a goal around that. It is a more positive way of looking at the tool (Rayann).

Nancy preferred the alternative score system because of the strength-based approach:

> The idea of quality is that I am going to look at your strengths. If you have strengths on conversation and I am going to try to bring up the areas that are weak. In looking at the report and we are in coaching now, we are going to see what can be changed and improved (Nancy).

Using the alternative scoring system also reflected the real intention of program quality assessment and improvement. Which helped promote a growth mindset and partnership rather than fixated with the numbers. As a coach shared:

> I look at it all. The ECERS-R give you more explanation on why the score low, the writing part. But I am always highlighting the staff that they are good at or otherwise they are going to be defeated. If they look at the good score, you can tell them how great they are. And then we look at the lower part and I said do not worry about that we will work on that. But no matter what, they always look at the low scores (Vicky).

To summarize, coaches seemed to prefer the alternative scoring system with detailed notes as it provided more information about the programs' strengths and areas of improvement. They reminded us that the assessment should capture the whole program and serve for the goal of program quality assessment and improvement. Moreover, the alternative scoring system helped cultivate a strength-based collaborative partnership as providers' strengths were acknowledge and recognized.

Theme 3: Data-based goal setting was grounded in valid information from the tools, coaches' knowledge about the tools and the programs, and the ability to interpret and negotiate with program providers, which could vary across coaches. Many coaches or providers might not believe that the information assessed by the tools was valid and would not solely rely on the number to make decisions. Instead, they also utilized their knowledge about the tools and the programs. And coaches needed to interpret the results carefully and set goals collaboratively with the providers. As for goal setting, three style seem to emerged: some preferred to rely on the tools and "teach to the test," while others used an empowerment

approach and encourage providers to "go with their heart." And some might choose a more balanced approach.

Coaches' personal belief about the tools might influence to what extent they considered the information as valid and reliable. Despite the validity evidence about the tools that was available, coaches might still question to what extent the information was valid and credible. For example, Anne did not believe that using TV was appropriate:

> I was very involved in a state screen time reduction program of getting kids off the screen. And they have changed in the book so that it is more official to use the TV and I think it is really sending a wrong message. I think they should get a full 7 point if they do not use any screen time. I would like to reward people for turning off the TV. And there is plenty of research back that up. If you do not use the TV, it is not applicable. So it doesn't count. But if you do use TV, you can get 7 points. They should get 7 points for not using any screen time (Anne).

As they worked with program providers, many coaches also tried hard to explain to the program providers that the assessment was just a snapshot and it might not truly reflect the program quality. As Vicky shared:

> Sometimes the provider gets very upset on what is written and do not think that is what actually happened. The provider argued that they only saw a small session of the program. That is the main complaint. And I whispered to them, "this is just a snapshot" (Vicky).

And their experience suggested that data from different tools might also appear to be contradicted. Take language and communication for example, Rayann shared that:

> When we got the results back and see some sites get a score of 3 on CLASS tool then the score they get in the language and reasoning was a 1 in that section. And they were being observed by the CLASS and ECERS assessors at the same time on that day. So we are like how is that work? That doesn't make any sense. They say it is not subjective but I mean it is human being seeing and observing and interpreting the behavior. And to me, that is still very subjective (Rayann).

In these circumstances, coaches not only needed to draw on their knowledge about the tool, but also be able to interpret the results and explain well to the providers, which could be quite challenging. As Tara pointed out:

> If they score high on the CLASS tool for language modeling, feedback loop, that kinds of thing, it confuses me when they score low on the ECERS-R or FCERS. I know they kind of measure different things but the ECERS-R tool is more about monitoring the frequency and the CLASS measures quality. But it is hard to explain to the providers why you got such a high score over here when they see the language. And then when they look at the other one and they got a low score. And try to explain to them is a little bit tricky for me because they see language as one thing (Tara).

With the available information and working knowledge, coaches also needed to prioritize and negotiate with program providers to set meaningful goals. As shown below, the negotiation process could vary considerably.

1.Teach to the test. Some might take on a more restricted approach and focus on getting the scores needed. Take Anne for example,

> For now, I have to teach to the test. People set up a system where this is the only way to get the score. I have many many conversations with the providers and if you want the

> score, this is the way you need to do. Some of them would say ok I would do that and some of them say screw it I do not want the point. For example, a Montessori site is not going to bring in dramatic play. Otherwise, they will because they want the point (Anne).

2. Balanced approach. Other coaches might try to balance the requirement and choosing meaningful objectives. Below was Benny's way of goal setting negotiation:

> I think the ECERS is a great wonderful tool but the first thing I realize is that this tool is not like ABC… This tool does not have all the answers. If you really read through the items, some of them actually conflict with each other. So it is about choosing. It is about passing but it is about choosing. You do not have to get everything right to pass, right? So when I work with a program, I first say let's look at your mission and your goals, and what you want your program to look about. And then let's look and find the ones that you will do well on because this is what you want, such as a Montessori school. Montessori schools have pieces that totally conflict with the tool, like block and dramatic play. So what they have to do then is going and find the things in this tool that they do really well there. And others that not going to score well are fine because you do not have to get it all. And then people aren't intimidated by the tool because then the tool serves its purpose for them to reach the goals that they want to reach (Benny).

Similar to Benny, Yan would intentionally balance the goals:

> Teachers always want to fix their environment first. What do I need to fix in the block area and what do I need to fix in the dramatic play? The area that they tend to shy away from is the language and the child interaction. It is easier to fix the environment than it is for you to fix your intentional teaching and what you supply for the kids. So for me I take it we are going to change something in your environment but at the same time, we will be

> working on some of your instructional support, and emotional support, your interaction and what you supply for the children (Yan).

3. Go with the heart. Some coaches prefer to empower the program providers and let them "go with their heart":

> I learned over the years on how to approach them. This is what ITERS and ECERS-R are saying and you have your own philosophy. You have something you truly believe what it should be. I can always tell you what the tool tells us. If we can figure out how to fit that in, that is great. But I can't really change your mind if you do not take it. I have a family child care person who is very wonderful and she has a gorgeous gorgeous room. She has implemented a lot of things in the room but there is always a fine line of having enough space that she is supposed to have. You go with your heart. I always tell them there are things that we can duplicate like fine motor and math. But you have to make the judgment call and decide what you think you want to work with. Some people came to get the level 4 but some people are happy with just a 3 again (Vicky).

In summary, data-based decision making was a complex process that was grounded in coaches' perception of the validity inferences at the item and scoring levels. In addition, the coaches need to draw on their knowledge about the tools and the programs to interpret the information. The negotiation process could vary considerably, ranging from score-focus, balanced, and empowerment approaches.

Chapter Five: Discussion

Despite over 30 states have adopted ECERS-R as part of their program quality assessment systems (Tout et al., 2015), the validity of ECERS-R remains as a major concern. One of the most pressing question centers upon decisions users must make between the two scoring systems. This question has implications not just for researchers interested in the psychometric properties of assessments, but also for coaches who use the ECERS-R for training, coaching, or technical assistance in their state's QRIS. The purpose of this study, therefore, was to compare the differences of validity evidence based on the two scoring systems in the context of a state's QRIS. While previous studies mainly focus on examining predicted validity, no study systematically examines the difference of validity evidence based on its two scoring systems. This study is the first attempt to deepen our understanding of how scoring systems may affect advanced validity evidence. This study also provides original consequential validity insights from coaches on how the scoring systems may influence their decision making and continuous quality improvement. Below I highlight five important findings and discuss them with relevant literature and policy context.

First, findings suggested that the quality scores between the two scoring systems could be dramatically different. We observe statistical mean differences between the two scoring systems in both total score and subscale score. The ECERS-R alternative score was almost 33 points higher than the stop score and this pattern was observed at the subscale level as well. For example, the mean stop score for Personal Care was only 15.42 (*SD*=8.24, ranging from 2.78 to 41.67), while the alternative score was 74.83 (*SD*=8.63, ranging from 50.24 to 90.28). This finding suggests that the priority can be totally different depending on which scoring system is used. For example, many programs may be low in personal care domain with the stop score

system and thus prioritizes that as their program improvement goal. But if the alternative scoring system is introduced, the score in personal care domain can be much higher compared to other domains. assessment developers should reconsider the scoring system based on my research findings.

Preliminary research (Hofer, 2008, 2010) indicates that 25% of the childcare centers can move above one state's cutoff for more funding using alternative scoring instead of stop scoring methods. If the assumption of stop scoring is not held (Gordon et al., 2015), the use of stop scoring system might introduce systematic measurement error at the scoring level. This finding could have important implications for program quality improvement. As findings in the qualitative component suggest, some coaches may rely on the ECERS-R score to prioritize goals. For classroom assessments, the goal should focus on providing specific information about program strengths and areas of improvement to support coaches' program quality decision (e.g., De Grosso et al., 2010; Cusumano et al., 2006). Clear goals for the facets of quality and corresponding child outcomes addressed in an initiative provide an essential foundation for a quality improvement initiative (Aiken & Akers, 2011; Tout et al., 2010). However, the priority can be totally different depends on which scoring system is used.

Second, findings suggest some ECERS-R subscales (e.g., space and furnishing) significantly predict children's language and early math outcomes. However, the effect sizes are small. This finding fails to support the developers' claim that the predictive validity of ECERS-R is well-established (Harms et al., 2015). The fact that both scoring systems had small relationship with child outcomes seems to reinforce the need to reconsider the items and factor structure.

This finding is consistent with previous studies (e.g., Burchinal et al., 2011; Gordon et al., 2013; Mayer & Beckh, 2016; Sylva et al., 2006). Based on factor and IRT analyses, Mayer and

Beckh (2016) provide evidence that the small relations between childcare quality and child outcomes may be in part due to the weak psychometric properties of the measure itself. Consequently, the associations between child care quality and child outcomes may be underestimated. The fact that both scoring systems had small relations with child outcomes seems to reinforce the need to reconsider the items and factor structure. Previous research (Gordon et al., 2015) indicates that over half of the indicators of the ECERS-R activities subscale did not fit together to define a single underlying dimension. And other reviewed literature (e.g., Cassidy et al., 2005; Clifford et al., 2005; Early et al., 2006; Frede et al., 2007; Sakai et al., 2003; Gordon et al., 2015; Gordon et al., 2013) suggest contradicted factor structures compared to the developers (Clifford et al., 2010).

Future assessment development should conceptualize the factor structure based on strong theory and inputs from experts to establish strong face validity. It is important to develop items or indicators that align with each factor. A focus on domain specificity can be beneficial from a measurement perspective, as psychometricians recommend that measure development begin by carefully defining each dimension, differentiating the dimensions from one another, and writing items specific to each dimension (Wolfe & Smith, 2007a, 2007b). The current stop-scoring approach challenges policymakers, coaches, and researchers who wish to isolate particular quality components in order to examine how they inter-correlate and how they relate to child outcomes. Further research is needed to check whether alternative scoring approaches would produce more domain-specific measures of quality than the standard stop-scoring approach. This line of work is beyond the ECERS-R to consider whether and how to develop measures of child care quality specific to domains of child development (Burchinal et al., 2011; Forry et al., 2009; Vandell & Wolfe, 2000; Zaslow et al., 2006; Zaslow et al., 2011).

Third, my findings suggest that coaches felt frustrated to comply with some quality indicators due to lack of feasibility, cultural adaptation, and compatibility with program philosophy and facility. Based on the qualitative findings, it is critical for local QRIS teams to review the scoring profiles across programs to identify item(s) that would be the priority and provide guidelines on items that are not feasible to implement. Moreover, the assessment developers should put more thought about integrating culturally responsive practices and inclusive practices into the protocol. As our field is serving more children from culturally and linguistically diverse backgrounds, program quality practices should reflect this change and integrate evidence-based practices for this population. Furthermore, it is critical for assessment developers to provide guidelines on how the indicators may be adapted to various cultural and linguistic contexts. For example, the item can describe what a practice may look like in general classroom settings and provides examples to illustrate how the practices may be acceptable within different cultural and linguistic context.

Previous research reports that ECERS-R falls short in key components of culturally sensitive practice, such as communicating with families in their home language. For example, Sakai et al. (2003) studied a diverse sample in California and found that programs fall short on linguistic match can still be rated as high in quality. This study adds to the fact that some quality indicators such as hand washing may warrant further consideration to make sure they are feasible to implement. Moreover, while it is helpful to clearly define and quantify the items, such description made coaches feel like it is too restricted and this is the only way to do things without considering the cultural and linguistic adaptation.

Fourth, most of the coach participants suggest that they prefer the alternative scoring system. The alternative scoring system helps them gain a more comprehensive picture about the

program strengths and areas of improvement. More importantly, the alternative scoring system helps cultivate a strength-based coaching partnership. From policy makers or researchers' perspectives, the alternative scoring system might be just a simple choice or preference. But from coaches' perspectives, it represents the idea of recognizing and reinforcing providers' good practices.

The underlying assumption of the current stop-scoring may contradict with the strength-based approach to quality assessment and improvement. The ECERS-R *stop-scoring* system postulates a specific order in item scores, since indicators of higher categories can only be rated if indicators of lower categories are fulfilled. Although the ECERS-3 encompasses some changes consisting of refining some indicators, it retains much of the structure of ECERS-R. To receive high scores on one item, all of the qualitative indicators of the lower categories need to be met (Harms et al., 2005).

Previous literature (Akers & Aikens, 2012; Isner et al., 2011) suggests that a standard sequence of activities be used in the approach that first emphasize relationship-building and goal setting in initial work with providers/program; then focus on implementation of an action plan with clear roles, provision of feedback and reflection; and finally provide an opportunity to assess the process and plan for sustainability. To make sure the goal is meaningful and encourage actual implementation, it is critical to building a trust relationship. A core feature of continuous quality improvement is using data to make decisions about quality improvement priorities, and on developing the capacity for ongoing assessment and improvements (Mitchell, 2012; Wiggins & Mathias, 2013). More importantly, it is critical to promote a culture of positive change and assessment of strengths and needs (Tout et al., 2015).

When redesigning the instrument, it is important to consider its implication at higher level

validity inferences. Ideally, subscales should be separated based on the factor structure and the scoring should not depend on each other. More importantly, the stop-scoring system may provide inaccurate information as the program quality trait is not fully represented (Kane, 2006). Concord with Gordon et al. (2015), I strongly encourage the assessment developer to continue to refine the items and scoring based on psychometric evidence and measurement theory. Meanwhile, the goal should focus on the interpretation of performance in context rather than just paying attention to scores (Kane, 2006).

Fifth, the findings suggest that coaches vary in their ability to interpret data from quality measures, and the ability to work with program providers to define quality improvement goals. These findings call for attention to provide systematic support to both coaches and program providers to cultivate a shared understanding of QRIS. The findings suggest that coaches not only need to draw on their knowledge about the tool, but also be able to interpret the results and explain well to the providers, which can be quite challenging. To overcome this challenge, it is critical to provide training and guidance based on best coaching practices. Local QRIS agencies may also consider using community of practices to encourage shared understanding.

Previous research suggests that coaches or consultant may use an environmental assessment tool to assess program needs and to inform goal development jointly (Cusumano et al., 2006; De Grosso et al., 2010; Wesley et al., 2010). However, data-based decision-making is a complex process grounded on coaches' perception of the validity inferences at the item and scoring levels. In addition, the coaches need to draw on their knowledge about the tools and the programs to interpret the information. The negotiation process could vary considerably, ranging from score-focus, balanced, and empowerment approaches.

Several limitations should be noted when interpreting the findings. First, the quantitative sample is relatively small and based on available data. The data set also contains limited demographic information about the children. Therefore, selection bias and generalizability may be affected. Second, the study relies on secondary data sources for a number of variables, which contained missing data points. At present, I treated the controlling variables like subsidy as missing at random and used multiple imputations to stimulate data based on the available variables. Because HLM requires complete data at the highest level of the model, sites that do not have complete data are excluded from analysis and further impacted the sample size. Third, some child outcome instruments may not have strong validity and may potentially introduce systematic errors. As described in the method section, some instruments only have reliability data but do not have much validity information. Fourth, when multiple analyses are conducted without sufficient statistical power, it is at risk to reject or accept the hypothesize when it is in fact due to random errors (i.e., type 1 and type 2 errors). I conducted a number of HLM analysis and the results may be biased due to type 1 or type 2 errors. Fifth, the weak predictive validity may have to do with the inappropriate factor structures. Due to insufficient sample size at the site level, I can not conduct factor analysis and use the results to provide construct validity evidence and support other analysis.

For the qualitative component, the reader should also bear in mind that that participants were recruited based on their qualification. The convenience sampling may limit to what extent their perspectives are represented at a broader scope. Although I recruited 13 participants from different geographical regions to maximize the findings' generalization, the fact that only one source of evidence might not truly reflect coaches' perceptions. Moreover, some of the feedback

about program providers are learned through the coaches. It is not known whether the findings fully represent their perspectives.

Despite the limitations, this study is the first attempt to deeper our understanding of how scoring systems may affect advanced validity evidence (e.g., extrapolation, interpretation, implication). It also provides original consequential validity insights from coaches on how the scoring systems may influence their decision making and continuous quality improvement. Future research may apply this method with a different sample to replicate the findings. Also, it would be interesting to compare the two scoring systems as the ECERS-3 is implemented. Moreover, more research should be conducted to further review the items (i.e., face validity) and structure the scale based on factor analysis and other psychometric evidence. It will be also important to continue to gather practitioner feedback to inform assessment development. Last but not least, it is important to examine effective practices to promote joint assessment and data-based decision making.

References

Aboud, F.E., & Hossain, K. (2011). The impact of preprimary school on primary school achievement in Bangladesh. *Early Childhood Research Quarterly*, *26* (2), 237-246. doi: 10.1016/j.ecresq.2010.07.001

Achenbach, T.M., & Rescorla, L.A. (2000). *Manual for the ASEBA preschool forms and profiles.* Burlington, VT: University of Vermont Department of Psychiatry. ISBN 0-938565-68-0

Adams, G., Tout, K., & Zaslow, M. (2007). *Early care and education for children in low-income families: Patterns of use, quality, and potential policy implications.* Prepared for the Urban Institute and Child Trends Roundtable on Children in Low-Income Families. Washington, DC.

Aikens, N. & Akers, L. (2011). *Background review of existing literature on coaching*. Washington DC: Mathematica Policy Research.

Anders, Y., Rossbach, H.-G., Weinert, S., Ebert, S., Kuger, S., Lehrl, S. and von Maurice, J. (2012). Home and preschool learning environments and their relations to the development of early numeracy skills. *Early Childhood Research Quarterly, 27,* 231–244. doi: 10.1016/j.ecresq.2011.08.003

Barnard, W., Smith, W. E., Fiene, R., & Swanson, K. (2006). *Evaluation of Pennsylvania's Keystone STARS quality rating system in child care settings.* Pittsburgh, PA: University of Pittsburgh School of Education, Office of Child Development.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2-3), 62-87.

Biersteker, L., Dawes, A., Hendricks, L., & Tredoux, C. (2016). Center-based early childhood care and education program quality: A South African study. *Early Childhood Research Quarterly*, *36*, 334-344. doi: 10.1016/j.ecresq.2016.01.004

Boller, K., Blair, R., Del Grosso, P., & Paulsell, D. (2010). *Better beginnings: The seeds to success modified field test: Impact evaluation findings.* Princeton, NJ: Mathematica Policy Research.

Boller, K., Tarrant, K. & Schaack, D.D. (2014). *Early care and education quality improvement: A typology of intervention approaches.* OPRE Research Report #2014-36. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Bryant, D. M., Clifford, R. M., & Peisner, E. S. (1991). Best practices for beginners: Developmental appropriateness in kindergarten. *American Educational Research Journal*, *28*, 783–803. doi:10.3102/00028312028004783

Burchinal, M. R., Howes, C., & Cryer, D. (1997). The prediction of process quality from structural features of child care. *Early Childhood Research Quarterly*, *12*, 281–303. doi: 10.1016/s0885-2006(97)90004-1

Burchinal, P., Kainz, K., Cai, K., Tout, K., Zaslow, M., Martinez-Beck, I. & Rathgeb, C. (2009). *Early care and education quality and child outcomes, OPRE Policy Brief.* Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Burchinal, M. R., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. In M. Zaslow, I. Martinez-Beck,K. Tout, & T. Halle

(Eds.), *Quality measurement in early childhood settings* (pp.11–31). Baltimore, MD: Brookes.

Cassidy, D. J., Hestenes, L. L., Hegde, A., Hestenes, S., & Mims, S. (2005). Measurement of quality in preschool child care classrooms: An exploratory and confirmatory factor analysis of the early childhood environment rating scale–revised. *Early Childhood Research Quarterly, 20*, 345–360. doi: 10.1016/j.ecresq.2005.07.005

Clauser, B.E., Harik, P., & Clyman, S.G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. *Journal of Educational Measurement, 37*, 245-261. doi: 10.1111/j.1745-3984.2000.tb01085.x

Charmaz, K. (2002). Qualitative interviewing and grounded theory analysis. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of interview research: Context & method* (pp. 675–694). London, England: SAGE.

Child Trends. (2014). *QRIS compendium*. Retrieved from http://qriscompendium.org/

Clifford, R. M., Sideris, J., & Neitzel, J. (2012, June). *New scoring mechanisms for the ECERS-R*. Paper presented at the National Association for the Education of Young Children's 21st National Institute for Early Childhood Professional Development, Indianapolis, IN.

Clifford, R. M., Barbarin, O., Chang, F., Early, D., Bryant, D., Howes, C., & Pianta, R. (2005). What is pre-kindergarten? Characteristics of public pre-kindergarten programs. *Applied Developmental Science, 9,* 126–143. doi: 10.1207/s1532480xads0903_1

Clifford, R. M., Reszka, S. S., Rossbach, H. (2010). *Reliability and validity of the Early Childhood Environment Rating Scale.* Retrieved from ers.fpg.unc.edu/sites/ers.fpg.unc.edu/files/ReliabilityEcers.pdf.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. doi: 10.1037/0033-2909.112.1.155

Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Copple, C. & Bredekamp, S. (Eds.). (2009). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8*. Washington, DC: National Association for the Education of Young Children.

Crano, W. D., & Brewer, M. B. (2002). *Principles and methods of social research*. Mahwah, N.J: Lawrence Erlbaum Associates.

Cronbach, L. J. (1988). Internal consistency of tests: Analyses old and new. *Psychometrika, 53*, 63–70.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological Tests and Personnel Decisions.* University of Illinois Press, Urbana

Cryer, D. (1999). Defining and assessing early childhood program quality. *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: Kaplan.

Cusumano, D. L., Armstrong, K., Cohen, R., & Todd, M. (2006). Indirect impact: How early childhood educator training and coaching impacted the acquisition of literacy skills in preschool students. *Journal of Early Childhood Teacher Education, 27*(4), 363-377. doi: 10.1080/10901020600996166

Del Grosso, P., Hallgren, K., Paulsell, D., & Boller, K. (2010). *Better beginnings: The Seeds to Success Modified Field Test: Implementation lessons*. Princeton, NJ: Mathematica Policy Research.

Denny, J. H., Hallam, R., & Homer, K. (2012). A multi-instrument examination of preschool classroom quality and the relationship between program, classroom, and teacher characteristics. *Early Education & Development*, *23*(5), 678-696. doi: 10.1080/10409289.2011.588041

Dunn, L. (1993). Proximal and distal features of day care quality and children's development. *Early Childhood Research Quarterly*, *8*(2), 167-192. doi:10.1016/S0885-2006(05)800894

Dunn, L., & Dunn, D. (2007). *Peabody Picture Vocabulary Test, Fourth Edition* (PPVT-4). Minneapolis, MN: NCS Pearson Psychopath.

Dunn, D. M., & Dunn, L. M. (2007). *Peabody picture vocabulary test*, *Manual (fourth edition)*. Minneapolis, MN: NCS Pearson Inc.

Forry, N., Vick, J., & Halle, T. (2009). *Evaluating, developing, and enhancing domain-specific measures of child care quality* (OPRE Research-to-Policy Brief No. 2). Washington, DC: Administration for Children and Families, Office of Planning, Research, and Evaluation.

Frank Porter Graham Child Development Institute. (1999). *Early developments*. Chapel Hill: University of North Carolina at Chapel Hill.

Frede, E., Jung, K., Barnett, W. S., Lamy, C. E., & Figueras, A. (2007). *The abbott preschool program longitudinal effects study (APPLES). Interim report.* New Brunswick, NJ: National Institute for Early Education Research.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago, IL: Aldine Press.

Greenfield, D. B., Jirout, J., Dominguez, X., Greenberg, A.C., Maier, M., & Fuccillo, J. (2009). Science in the preschool classroom: A programmatic research agenda to improve science readiness. *Early Education and Development, 20*(2), 238-264. doi:

10.1080/10409280802595441

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS–R with implications for measures of child care quality and relations to child development. *Developmental Psychology, 49*(1), 146–160. doi: 10.1037/a0027899

Gordon, R. A., Hofer, K. G., Fujimoto, K., Risk, N., Kaestner, R., & Korenman, S. (2015). Identifying high-quality preschool programs: New evidence on the validity of the early childhood environment rating scale–revised (ECERS–R) in relation to school readiness goals. *Early Education and Development, 1,* 213-234, doi: 10.1080/10409289.2015.1036348

Harms, T., & Clifford, R. M. (1980). *Early childhood environment rating scale.* New York, NY: Teachers College Press.

Harms, T., & Clifford, R. M. (1982). Assessing preschool environments with the early childhood environment rating scale. *Studies in Educational Evaluation*, *8*(3), 261-269.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale (revised edition)*. New York, NY: Teachers College Press.

Harms, T., Clifford, R. M., & Cryer, D. (2005). *Early childhood environment rating scale* (Rev. ed.). New York: Teachers College Press.

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. Pianta, M. Cox, & K. Snow (Eds.), *School readiness and the transition to kindergarten in the era of accountability* (pp. 49–84). Baltimore: Brookes.

Harms, T., Clifford, R. M., & Cryer, D. (2015). *Early childhood environment rating scale (third edition)*. New York, NY: Teachers College Press.

Helburn, S., Culkin, M. I., Morris, J., Mocan, N., Howes, C., Phillipsen, L., Bryant, D., Clifford, R., Cryer, D., Peisner-Feinberg, E., Burchinal, M., Kagan, S. L., & Rustici, J. (1995). *Cost, quality, and child outcomes in child care centers, public report, (2nd ed.).* Denver: Economics Department, University of Colorado at Denver.

Herrera, M. O., Mathiesen, M. E., Merino, J. M., & Recart, I. (2005). Learning contexts for young children in Chile: process quality assessment in preschool centers. *International Journal of Early Years Education*, *13*(1), 13-27.

Hofer, K. G. (2008). Measuring quality in prekindergarten classrooms: Assessing the Early Childhood Environment Rating Scale. Unpublished dissertation, Vanderbilt University, Nashville, TN.

Hofer, K. G. (2010). How measurement characteristics can affect ECERS-R scores and program funding. *Contemporary Issues in Early Childhood*, *11*(2), 175–191. doi: 10.2304/ciec.2010.11.2.175

Hooks, L. M., Scott-Little, C., Marshall, B. J., & Brown, G. (2006). Accountability for quality: One state's experience in improving practice. *Early Childhood Education Journal*, *33*(6), 399-403. doi: 10.1007/s10643-006-0065-3

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly, 23*, 27–50.doi: 10.1016/j.ecresq.2007.05.002

Howes, C., & Smith, E. W. (1995). Relations among child care quality, teacher behavior, children's play activities, emotional security, and cognitive activity in child care. *Early Childhood Research Quarterly*, *10*(4), 381-404. doi: 10.1016/0885-2006(95)90013-6

Ishimine, K., Wilson, R., & Evans, D. (2010). Quality of Australian childcare and children's social skills. *International Journal of Early Years Education*, *18*(2), 159-175. doi: 10.1080/09669760.2010.494430

Isner, T., Tout, K., Zaslow, M., Soli, M., Quinn, K., Rothenberg, L., & Burkhauser, M. (2011). *Coaching in early care and education programs and quality rating and improvement systems (QRIS): Identifying promising features.* Washington, DC: Child Trends.

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action *Administrative Science Quarterly, 24*(4), 602-611.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.

La Paro K. M., Thomason, A. C., Lower, J. K., Kintner-Duffy, V. L., & Cassidy, D. J. (2012). Examining the definition and measurement of quality in early childhood education: A review of studies using the ECERS-R from 2003 to 2010. *Early Childhood Research & Practice, 14*(1). Retrieved June 18, 2014, from http://ecrp.uiuc.edu/v14n1/laparo.html

La Paro, K.M., Williamson, A.C., Hatfield, B. (2014). Assessing quality in toddler classrooms using the CLASS-Toddler and the ITERS-R, *Early Education and Development, 25*(6), 875–893.

Layzer, J. I. & Goodson, B. D. (2006). The quality of early care and education settings: Definitional and measurement issues. *Evaluation Review, 30 (5),* 1-21.

Li, K., Hu, B. Y., Pan, Y., Qin, J., & Fan, X. (2014). Chinese early childhood environment rating scale (trial)(CECERS): A validity study. *Early Childhood Research Quarterly*, *29*(3), 268-282.

Li, K., Pan, Y., Hu, B., Burchinal, M., De Marco, A., Fan, X., & Qin, J. (2016). Early childhood education quality and child outcomes in China: Evidence from Zhejiang Province. *Early Childhood Research Quarterly*, *36*, 427-438. doi: 10.1016/j.ecresq.2016.01.009

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D.,. . .& Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*, 732–749. doi: 10.1111/j.1467- 8624.2008.01154.x

Mayer, D., & Beckh, K. (2016). Examining the validity of the ECERS–R: Results from the German national study of child care in early childhood. *Early Childhood Research Quarterly*, *36*, 415-426. doi: 10.1016/j.ecresq.2016.01.001

Mayer, D., & Beckh, K. (2016). Examining the validity of the ECERS–R: Results from the German National Study of Child Care in Early Childhood. *Early Childhood Research Quarterly*, *36*, 415-426. doi: 10.1016/j.ecresq.2016.01.001

Merriam, S. (2009). *Qualitative research: A guide to design and implementation*. San Francisco, CA: Jossey-Bass.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan

Mitchell, A.W. (2005). *Stair steps to quality: A guide for states and communities developing quality rating systems for early care and education.* United Way Success by 6.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook.* Thousand Oaks, CA: SAGE.

Morse, J. M. (1991). *Qualitative nursing research: A contemporary dialogue*. Newbury Park, Sage.

National Center on Child Care Quality Improvement. (2013). *QRIS elements administration for children and families*: Office of Child Care. Available from: *https://childcareta.acf.hhs.gov/resource/qris-elements*

National Center on Child Care Quality Improvement. (2015). *QRIS standards, levels, and rating systems.* Administration for Children and Families: Office of Child Care. Available from: https://occqrisguide.icfwebservices.com/files/QRIS Levels Rating.pdf

National Child Care Information and Technical Assistance Center [NCCIC]. (2010). *Quality rating and improvement system resource guide.* Washington, DC: Administration for Children and Families, U.S. Department of Health and Human Services.

Palsha, S., Ritchie, S., Sparling, J., Maxwell, K., Crawford, G. and Lim, C. (2007). *A first school lens on instructional practices and curriculum: Changing schools to benefit early childhood professionals, young children, and their families*. Foundation for Child Development.

Pakarinen, E., Lerkkanen, M. K., Poikkeus, A. M., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., & Nurmi, J. E. (2010). A validation of the classroom assessment scoring system in Finnish kindergartens. *Early Education & Development, 21*(1):95-124.

Paulsell, D., Tout, K., & Maxwell, K. (2013). Evaluating implementation of quality rating and improvement systems. In T. Halle, A. Metz, and I. Martinez-Beck (Eds.), *Applying*

implementation science in early childhood programs and systems (pp. 269-293). Baltimore, MD: Brookes Publishing.

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. doi: /10.1016/j.ecresq.2004.07.006

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2007). *Classroom Assessment Scoring System.* Baltimore, MD: Brookes Publishing.

Ponitz, C. C., McClelland, M. M., Jewkes, A. M., Connor, C. M., Farris, C. L., & Morrison, F. J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Research Quarterly*, *23,* 141–158. doi: 10.1016/j.ecresq.2007.01.004.

Ponitz, C. C., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral regulation and its contributions to kindergarten outcomes. *Developmental Psychology, 45*, 605-619. doi: 10.1037/a0015365.

R Development Core Team (2008). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nded.)*. Newbury Park: Sage

Ryan, G.W., & Bernard, H.R. (2003) Techniques to Identify Themes. *Field Methods*, *15*(1):85–109. doi: 10.1177/1525822X02239569

Sarama, J., Clements, D.H., & Wolfe, C.B. (2010). *Tools for early mathematics assessment' instrument and manual.* Columbus, OH: McGraw-Hill.

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 editions. *Early Childhood Research Quality, 18*, 427–445.doi: 10.1016/j.ecresq.2003.09.004Schneider

Sabol, T. J., & Pianta, R. C. (2014). Do standard measures of preschool quality used in statewide policy predict school readiness? *Education*, *9*(2), 116-164. doi: 10.1162/EDFP_a_00127

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

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects: The High/Scope Perry Preschool Study through age 40*. Ypsilanti, MI: High/Scope Press.

Schaack, D., Le, V. N., & Setodji, C. M. (2013). Examining the factor structure of the Family Child Care Environment Rating Scale—Revised. *Early Childhood Research Quarterly*, *28*(4), 936-946. doi: 10.1016/j.ecresq.2013.01.002

SPSS Inc. (2006). *SPSS base 15.0 for Windows user's guide.* Chicago IL: SPSS Inc.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks, CA: Sage.

Soderberg, J., Joseph, E.G., Stull, S., Hassairi, N. (2016). *Early achievers standards validity study.* Childcare Quality and Early Learning Center for Research & Professional Development (CQEL).

Sylva, K., Siraj-Blatchford, I., Taggart, B., Sammons, P., Melhuish, E., Elliot, K., &Totsika, V. (2006). Capturing quality in early childhood through environmental rating scales. *Early Childhood Research Quarterly, 21*, 76–92. doi: 10.1016/j.ecresq.2006.01.003

Sylva, K., Siraj-Blatchford, I., Taggart, B., Sammons, P., Melhuish, E., Elliot, K., & Totsika, V. (2006). Capturing quality in early childhood through environmental rating scales. *Early Childhood Research Quarterly*, *21*(1), 76-92. doi: 10.1016/j.ecresq.2006.01.003

Tout, K., Epstein, D., Soli, M., & Lowe, C. (2015). *A blueprint for early care and education quality improvement initiatives.* Publication# 2015-07. *Child Trends*.

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., & Boller, K. (2010). *ACF-OPRE report: Compendium of quality rating systems and evaluations*. Washington, DC: Administration for Children and Families, Office of Planning, Research, and Evaluation.

Tout, K., & Sherman, J. (2005). *Inside the preschool classroom: A snapshot of quality in Minnesota's child care centers.* Child Trends and the Minnesota Child Care Policy Research Partnership. St. Paul, MN: Minnesota Department of Human Services.

U.S. Department of Health and Human Services (2003, May). *Head Start FACES 2000: A whole-child perspective on program performance, Fourth Progress Report.* Washington DC: Administration for Children and Families, Department of Health and Human Services.

Vandell, D. L., & Wolfe, B. (2000). *Child care quality: Does it matter and does it need to be improved?* Washington, DC: U.S. Department of Health and Human Services.

Warash, B. G., Markstrom, C. A., & Lucci, B. (2005). The Early Childhood Environment Rating Scale-Revised as a tool to improve child care centers. *Education*, *126*(2), 240.

Weiland, C., Ulvestad, K., Sachs, J., & Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program. *Early Childhood Research Quarterly*, *28*(2), 199-209. doi: 10.1016/j.ecresq.2012.12.002

Wen, X., Bulotsky-Shearer, R. J., Hahs-Vaughn, D. L., & Korfmacher, J. (2012). Head Start program quality: Examination of classroom quality and parent involvement in predicting children's vocabulary, literacy, and mathematics achievement trajectories. *Early Childhood Research Quarterly*, *27*(4), 640-653. doi: 10.1016/j.ecresq.2012.01.004

Wilcox-Herzog, A., McLaren, M., Ward, S., & Wong, E. (2013). Results from the quality early childhood training program. *Journal of Early Childhood Teacher Education*, *34*(4), 335-349. doi: 10.1080/10901027.2013.845635

Wiggins, K. & Mathias, D. (2013). *Continuous quality improvement: An overview report for state QRIS leaders.* BUILD Initiative.

Wolfe, E. W., & Smith, E. V., Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: part II—validation activities. In E.V. Smith Jr., & R. M. Smith (Eds.), *Rasch Measurement: Advanced and Specialized Applications* (pp. 243–290). Maple Grove, MN: JAM Press.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock- Johnson III.* Rolling Meadows, IL: Riverside.

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M., Espinosa, L., Gormley, W., Ludwig, J.O., Magnuson, K.A., Phillips, D.A., & Zaslow, M.J. (2013). *Investing in our future: The evidence base on preschool education.* New York: Foundation for Child Development and Ann Arbor, MI: Society for Research in Child Development.

Zaslow, M., Halle, T., Martin, L., Cabrera, N., Calkins, J., Pitzer, L., & Margie, N. G. (2006). Child outcome measures in the study of child care quality. *Evaluation Review*, *30*, 577–610. doi: 10.1177/0193841X06291529

Zaslow, M., Tout, K., & Martinez-Beck, I. (2010). *Measuring the Quality of Early Care and Education Programs at the Intersection of Research, Policy, and Practice,* OPRE Research-to-Policy, Research-to-Practice Brief OPRE 2011-10a. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Zaslow, M., Tout, K., Halle, T., Vick, J., & Lavelle, B. (2010). *Towards the identification of features of effective professional development for early childhood educators: A review of the literature.* Submitted to U.S. Department of Education. Washington, DC, Child Trends.

Zaslow, M. Martinez-Beck, I. Tout, K. & Halle, T. (2011). *Quality measurement in early childhood settings*. Baltimore, MD: Brookes.

Zellman, G. L., & Perlman, M. (2008). *Child-care Quality Rating and Improvement Systems in five pioneer states: Implementation issues and lessons learned*. Santa Monica, CA: RAND.

Zellman, G. L., & Karoly, L. A. (2012). *Incorporating child assessments into state early childhood quality improvement initiatives.* Santa Monica, CA: RAND Corporation.

Tables and Figures

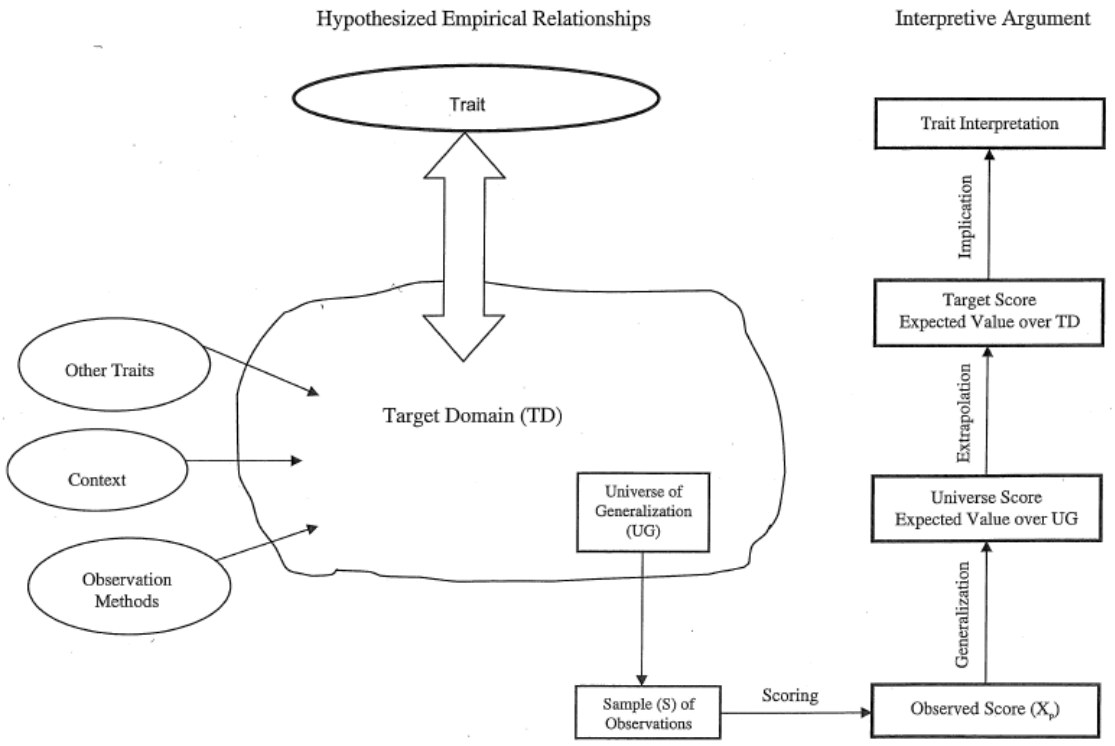

Figure 1. *Kane (2006)'s validity framework*

Table 1.

Validity Evidence by Reviewed Articles

		Scoring				Generalization	Extrapolation					Interpretation/	
		rubric appropriateness	consistency	accuracy	scaling	scale generalization	item face validity	criterion validity	convergent validity	factor structure	trait stability	predictive validity	social validity (decision; use)
First author	Year	expert judgment (survey; qualitative focus group)	interrater reliability test-retest internal consistency	independent scoring check	item response theory	sampling theory	expert judgment (survey; qualitative focus group)	correlation	correlation	factor analysis	measurement invariance	regression analysis	expert judgment (survey; qualitative focus group) consumer/practitioner feedback
Harms	1983		X				X						
Dunn	1993												
Scarr	1994					X							
Howes	1995											X	
Cryer	1999										X		
Sakai	2003						X	X		X			
Warash	2004											X	
Perlman	2004					X				X			
La Paro	2004							X					
Cassidy	2005						X					X	
Herrera	2005											X	
Cassidy	2005									X			
Hooks	2006											X	
Sylva	2006											X	
Ishimine	2010											X	
Aboud	2010											X	
Hofer	2010	X										X	
Sylva	2011											X	
Anders	2011											X	
Wen	2012											X	
Denny	2012							X					
Wilcox-herzog	2013											X	
Weiland	2013											X	
Gordon	2013				X					X		X	
Schaack	2013					X				X			
Soderberg	2014											X	
Sabol	2014											X	
Li	2014		X							X		X	
Gorden	2015				X		X						
Li	2016											X	
Mayer	2016				X			X		X		X	
Biersteker	2016											X	

Table 2

Children Characteristics

Characteristics	*N=531*	*%*	*M*	*Min*	*Max*	*SD*
Age			4.3	3	5.5	0.6
Gender						
Male	272	51.2				
Female	259	48.8				
Primary language						
English	411	77.4				
Other language	119	22.4				
Missing	1	0.2				
Race						
White	258	48.6				
Minority	170	32				
Missing	103	19.4				
Ethnicity						
Latino/a	92	17.3				
Other	336	63.3				
Missing	103	19.4				

Table 3

Children's Family Characteristics

Characteristics	*N=531*	*%*
Subsidy		
Yes	190	35.8
No	341	64.2
Parent education		
Less than 9th grade	53	10
Some high school	15	2.8
GED	8	1.5
High School Diploma	49	9.2
Some College	84	15.8
Associate's Degree	60	11.3
Bachelor's Degree	122	23
Masters' Degree or higher	140	26.4
Income		
$10,000 or less,	36	6.8
$11,000-$20,000	47	8.9
$21,000-$30,000	76	14.3
$31,000-$40,000	47	8.9
$41,000-$50,000	49	9.2
$51,000-60000	19	3.6
$61,000-$70,000	30	5.6
$71,000-$80,000	27	5.1
$81,000 or more	200	37.7

Table 4

Quantitative Instruments

Construct	Sign	Name	Citation
Program quality	ECERS-R	The Early Childhood Environment Rating Scale—Revised	Harms, Clifford, & Cryer, 1998
Program quality	CLASS	Classroom Assessment Scoring System Pre-K version	Pianta, La Paro, & Hamre, 2008
Early reading	WJ III	Woodcock–Johnson III Tests of Achievement	Woodcock, McGrew, & Mather, 2001
Early science	LENS	The Lens on Science	Greenfield et al., 2009
Early writing	EWA	The Early Writing Assessment	Adaptation of Puranik, 2011, 2012
Executive function	HTKS	Head Toes Knees and Shoulders	Ponitz et al., 2008
Early math	TEAM	Tools for Early Assessment in Math	Sarama, Clements, & Wolfe, 2010
Receptive language	PPVT-4	The Peabody Picture Vocabulary Test, Fourth edition	Dunn & Dunn, 2007
Social-emotional	CBCL	The Child Behavior Checklist	Achenbach & Rescorla, 2000

Table 5
Coach Participants

Coach	Age	Ethnicity	Experience in ECE	Experience as an Early Achiever coach?	Time teaching in an early childhood classroom	Region
Amy	50-59	White	>10 years	1-3 years	>10 years	Eastern
Anne	50-59	Latino/a	>10 years	1-3 years	1-3 years	Eastern
Benny	30-39	White	5-10 years	1-3 years	5-10 years	Eastern
Carta	>60	White	>10 years	3-5 years	>10 years	South
Daisy	30-39	White	>10 years	1-3 years	5-10 years	Northwest
Janet	50-59	White	>10 years	1-3 years	>10 years	Eastern
Jane	40-49	White	>10 years	1-3 years	>10 years	Southwest
Katy	30-39	Asian	1-3 years	1-3 years	1-3 years	South
Nancy	30-39	African American	>10 years	1-3 years	5-10 years	South
Rayann	40-49	White	>10 years	3-5 years	5-10 years	Central
Tara	40-49	Latino/a	>10 years	1-3 years	3-5 years	Eastern
Vicky	30-39	White	>10 years	<1 year	5-10 years	Eastern
Yan	50-59	Native Hawaiian	5-10 years	1-3 years	5-10 years	South

Table 6

Descriptive Statistics of Program Quality Measure

	Stop Score				Alternative Score			
Subscales/Items	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Mean ERS	46.31	11.36	23.41	73.82	79.16	7.69	56.59	90.77
Space and Furnishings	40.52	10.43	20.83	77.08	76.79	7.50	57.17	94.44
1. Indoor space	44.26	24.87	0.00	100.00	85.33	9.77	61.54	100.00
2. Furniture for routine care	31.70	27.57	0.00	100.00	80.37	11.20	37.50	100.00
3. Furniture for relaxation	49.09	26.89	16.67	100.00	75.60	15.29	44.44	100.00
4.Room arrangement	46.89	25.12	16.67	100.00	78.04	14.07	45.45	100.00
5. Space for privacy	50.78	33.63	0.00	100.00	77.21	19.62	42.86	100.00
6. Child-related display	54.90	23.44	16.67	100.00	72.46	16.61	33.33	100.00
7. Space for gross motor	14.85	8.18	0.00	50.00	69.76	15.25	30.00	90.00
8. Gross motor equipment	31.67	23.94	0.00	100.00	75.58	13.76	40.00	100.00
Personal Care Routines	15.42	8.24	2.78	41.67	74.83	8.63	50.24	90.28
9. Greeting/departing	33.15	31.60	0.00	100.00	79.49	13.15	50.00	100.00
10. Meals/snacks	13.69	6.40	0.00	16.67	84.59	11.55	33.33	94.44
11. Nap/rest	24.34	33.53	0.00	100.00	77.82	17.37	0.00	100.00
12. Toileting/diapering	5.27	7.76	0.00	16.67	72.33	13.61	35.71	92.86
13. Health practices	15.95	3.40	0.00	16.67	72.72	11.07	45.45	90.91
14. Safety practices	2.23	5.68	0.00	16.67	61.86	16.92	20.00	90.00
Language-Reasoning	52.13	16.28	16.67	91.67	76.96	10.29	47.41	95.46
15. Books and pictures	54.86	22.87	0.00	100.00	86.29	10.10	45.45	100.00
16.Encouraging communicate	57.88	23.77	0.00	100.00	81.30	12.72	44.44	100.00
17. Language reasoning	36.66	22.83	0.00	100.00	59.44	18.72	25.00	100.00
18. Informal use of language	59.13	22.17	16.67	100.00	80.81	15.46	36.36	100.00
Activities	52.52	14.16	26.67	94.44	77.91	9.17	55.67	97.22
19. Fine motor	67.23	26.38	16.67	100.00	86.55	11.32	55.56	100.00
20. Art	55.71	23.79	16.67	100.00	85.49	10.63	50.00	100.00
21. Music/movement	47.30	28.87	0.00	100.00	77.65	16.53	40.00	100.00
22. Blocks	39.77	25.74	0.00	100.00	66.21	18.85	18.18	100.00
23. Sand/water	48.15	22.90	0.00	100.00	68.30	20.21	0.00	100.00
24. Dramatic play	64.69	22.93	16.67	100.00	80.86	13.76	33.33	100.00
25. Nature/science	53.86	29.89	0.00	100.00	79.34	17.13	30.00	100.00
26. Math/number	49.87	29.48	0.00	100.00	76.63	17.23	30.00	100.00
27. Use of TV	20.72	21.63	0.00	100.00	67.82	19.74	10.00	91.00
28. Diversity	55.77	27.01	16.67	100.00	83.30	12.82	50.00	100.00
Interaction	56.35	23.30	6.67	100.00	82.59	12.82	45.94	100.00
29. Supervision of gross motor	41.65	31.99	0.00	100.00	74.22	18.90	10.00	100.00
30. General supervision	44.04	27.93	0.00	100.00	77.15	16.75	27.00	100.00
31. Discipline	53.52	32.65	0.00	100.00	81.54	15.06	33.33	100.00
32. Staff-child interaction	70.43	33.13	0.00	100.00	89.49	14.84	30.00	100.00

33. Child interactions	72.13	32.99	0.00	100.00	90.56	12.13	50.00	100.00
Program Structure	60.94	19.63	5.56	100.00	85.89	11.42	45.15	100.00
34. Schedule	52.70	28.52	0.00	100.00	82.90	13.80	36.00	100.00
35. Free play	55.05	27.87	0.00	100.00	84.75	12.73	30.00	100.00
36. Group time	75.55	27.47	0.00	100.00	91.60	11.36	50.00	100.00
37. Provisions for children with disabilities	56.37	36.10	0.00	100.00	79.31	22.82	14.00	100.00

Table 7

Paired T-test between Alternative and Stop Scorning

	M	*SD*	*SE*	*Lower*	*Upper*	*t*	*p*
SF	36.57	7.08	0.82	34.93	38.21	44.46	<.001
PC	59.06	7.00	0.81	57.44	60.68	72.57	<.001
LR	24.71	8.20	0.95	22.82	26.61	25.93	<.001
AC	24.94	8.06	0.94	23.08	26.81	26.63	<.001
IN	26.23	13.33	1.55	23.14	29.32	16.92	<.001
PS	24.77	11.27	1.31	22.15	27.38	18.91	<.001
ERS	32.71	5.54	0.64	31.43	34.00	50.79	<.001

Note. SF=space &furnishing; PC=personal care; LR=language reasoning; AC=activity; IN=instruction; PS=program structure; ERS= Environmental Rating Scale; df=119

Table 8

Comparison of Convergent Validity by the Two Scoring Systems

	ES_CLASS	CO_CLASS	IS_CLASS	CLASS
SF_alternative	.346	.303	.409	.406
PC_alternative	.416	.447	.161	.398
LR_alternative	.485	.506	.286	.495
AC_alternative	.224	.138	.149	.194
IN_alternative	.479	.563	.187	.479
PS_alternative	.472	.440	.275	.458
ERS_alternative	.537	.539	.310	.537
SF_stop	.302	.333	.276	.353
PC_stop	.287	.295	.049**	.245
LR_stop	.487	.525	.269	.497
AC_stop	.255	.171	.215	.245
IN_stop	.494	.587	.168	.487
PS_stop	.383	.382	.306	.414
ERS_stop	.529	.558	.303	.539

Note. All the relations are significant at .01 level, except one number with **; *N*=120

Table 9

Zero-Order Correlation by the Two Scoring Systems

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	SF_stop	1													
2	PC_stop	.374**	1												
3	LR_stop	.459**	.403**	1											
4	AC_stop	.375**	.328**	.394**	1										
5	IN_stop	.473**	.462**	.613**	.294**	1									
6	PS_stop	.370**	.320**	.496**	.453**	.534**	1								
7	SF_alternative	.752**	.203**	.458**	.445**	.448**	.370**	1							
8	PC_alternative	.412**	.665**	.431**	.180**	.585**	.327**	.369**	1						
9	LR_alternative	.420**	.383**	.899**	.384**	.670**	.530**	.448**	.525**	1					
10	AC_alternative	.354**	.218**	.496**	.858**	.285**	.436**	.556**	.200**	.521**	1				
11	IN_alternative	.415**	.363**	.680**	.285**	.882**	.554**	.477**	.663**	.787**	.387**	1			
12	PS_alternative	.319**	.320**	.553**	.456**	.509**	.865**	.350**	.397**	.586**	.485**	.612**	1		
13	ERS_stop	.654**	.593**	.792**	.630**	.831**	.778**	.601**	.578**	.787**	.600**	.790**	.737**	1	
14	ERS_alternative	.558**	.466**	.780**	.555**	.760**	.694**	.662**	.687**	.862**	.671**	.884**	.776**	**.903****	1

** Correlation is significant at the 0.01 level (2-tailed). *N*=120

Table 10

Descriptive Statistics of Outcome Measures

	N	*M*	*SD*	*Min*	*Max*
Pre PPVT	476	100.75	17.18	20	144
Post PPVT	399	103.20	14.61	52	136
Pre WJ	482	102.26	15.69	67	180
Post WJ	401	103.29	15.12	55	172
Pre TEAM	482	7.00	4.22	0	16
Post TEAM	393	8.89	3.86	0	17
Pre LENS	229	0.74	1.11	-2	3
Post LENS	308	1.27	1.10	-2	3
Pre CBCL	362	45.90	10.06	29	76
Post CBCL	224	46.60	10.08	29	79
Pre EWA Name	489	4.87	2.81	0	8
Post EWA Name	401	5.62	2.36	0	8
Pre EWA Word	489	6.08	4.21	0	18
Post EWA Word	397	7.61	4.75	0	18
Pre EF	472	10.92	15.66	0	60
Post EF	392	19.19	18.10	0	59
Pre TOQ	472	47.55	5.81	17	52
Post TOQ	393	48.99	4.31	26	52

Table 11

Zero-Order Correlation by the Two Scoring Systems and the Outcome Measures

	PPVT	WJ	TEAM	LENS	EWA_Name	EWA_Word	TOQ	EF	CBCL_TP
SF_stop	.094	**.113***	.076	-.005	.020	**.103***	.003	-.093	**-.149***
PC_stop	.040	.013	.059	**.130***	.075	.087	-.023	.010	-.020
LR_stop	.024	.003	.003	-.067	-.069	-.037	-.059	**-.118***	-.043
AC_stop	.025	.042	-.007	-.051	.025	-.011	-.008	-.089	**.137***
IN_stop	**.132****	**.148****	**.178****	**.128***	.092	**.126***	.022	.056	-.051
PS_stop	**.168****	.063	**.159****	**.181****	.063	.031	-.012	**.115***	.065
ERS_stop	**.126***	**.099***	**.128***	.089	.051	.069	-.014	-.008	-.007
SF_alternative	.096	.075	.007	-.032	-.025	.067	-.044	-.083	-.122
PC_alternative	-.017	-.004	.079	.068	.051	.059	.037	-.002	.011
LR_alternative	.037	.054	.068	-.004	.013	.019	-.007	-.060	.024
AC_alternative	.023	.032	-.041	-.043	-.036	-.046	.022	-.068	.069
IN_alternative	.068	.070	**.108***	.095	.031	.051	.016	.016	.018
PS_alternative	**.140****	.044	.081	**.123***	.016	-.005	.001	.081	.044
ERS_alternative	.080	.062	.076	.056	.014	.032	.007	-.018	.016

Note. * significant at .05 level; ** significant at .01 level; *N*=75

Table 12

HLM Fixed Effect Outcome Summary

		Alternative					Stop				
Outcome	Fixed Effect	*Coeff*	*SE*	*t*	*p*	*d**	*Coeff*	*SE*	*t*	*p*	*d**
PPVT	SF	0.11	0.04	2.55	**0.013**	0.30	0.09	0.05	1.94	0.057	0.23
	PC	0.01	0.05	0.29	0.771	0.03	0.01	0.05	0.11	0.911	0.01
	LR	0.04	0.05	0.80	0.428	0.09	0.02	0.05	0.52	0.608	0.06
	AC	0.01	0.05	0.20	0.840	0.02	-0.01	0.05	-0.14	0.893	-0.02
	IN	0.04	0.05	0.81	0.423	0.10	0.07	0.05	1.41	0.163	0.17
	PS	0.01	0.05	0.26	0.795	0.03	0.01	0.05	0.25	0.807	0.03
WJ	SF	0.05	0.03	1.69	0.096	0.20	0.03	0.03	0.84	0.406	0.10
	PC	-0.04	0.03	-1.17	0.246	-0.14	-0.06	0.03	-1.92	0.059	-0.23
	LR	0.02	0.03	0.45	0.652	0.05	0.02	0.03	0.61	0.545	0.07
	AC	-0.01	0.03	-0.32	0.750	-0.04	-0.03	0.03	-0.77	0.446	-0.09
	IN	0.03	0.03	0.72	0.473	0.08	0.04	0.03	1.10	0.274	0.13
	PS	0.00	0.03	0.06	0.949	0.01	-0.02	0.03	-0.53	0.598	-0.06
TEAM	SF	0.05	0.03	1.43	0.156	0.17	0.07	0.03	2.11	**0.039**	0.25
	PC	0.03	0.03	0.86	0.395	0.10	0.02	0.03	0.64	0.525	0.08
	LR	0.02	0.03	0.50	0.620	0.06	0.01	0.03	0.38	0.706	0.05
	AC	0.00	0.04	-0.10	0.923	-0.01	0.02	0.03	0.51	0.613	0.06
	IN	0.03	0.04	0.80	0.426	0.10	0.06	0.03	1.67	0.099	0.20
	PS	0.02	0.04	0.45	0.651	0.05	0.04	0.03	1.26	0.211	0.15
LENS	SF	-0.07	0.07	-0.95	0.347	-0.13	-0.07	0.07	-0.96	0.341	-0.13
	PC	0.07	0.06	1.12	0.268	0.15	0.03	0.06	0.50	0.619	0.07

	LR	-0.01	0.07	-0.10	0.917	-0.01	-0.09	0.07	-1.33	0.188	-0.18
	AC	-0.09	0.07	-1.27	0.208	-0.17	-0.06	0.07	-0.85	0.400	-0.11
	IN	0.04	0.07	0.59	0.558	0.08	0.04	0.07	0.53	0.601	0.07
	PS	0.14	0.08	1.85	0.070	0.25	0.14	0.07	1.95	0.057	0.26
Word	SF	-0.01	0.05	-0.19	0.847	-0.02	0.02	0.04	0.50	0.621	0.06
	PC	-0.04	0.05	-0.96	0.339	-0.11	0.02	0.04	0.49	0.623	0.06
	LR	0.04	0.05	0.80	0.427	0.09	0.03	0.05	0.55	0.584	0.06
	AC	0.02	0.05	0.33	0.739	0.04	0.06	0.05	1.30	0.199	0.15
	IN	0.00	0.05	0.00	0.998	0.00	0.04	0.05	0.85	0.401	0.10
	PS	0.06	0.05	1.15	0.253	0.14	0.09	0.05	2.07	**0.042**	0.24
Name	SF	0.28	0.22	1.24	0.218	0.15	0.38	0.22	1.71	0.092	0.20
	PC	0.13	0.23	0.56	0.575	0.07	0.32	0.22	1.48	0.144	0.17
	LR	0.09	0.23	0.37	0.714	0.04	0.01	0.23	0.03	0.976	0.00
	AC	0.01	0.24	0.06	0.950	0.01	0.16	0.23	0.71	0.479	0.08
	IN	0.06	0.24	0.27	0.789	0.03	0.31	0.23	1.37	0.175	0.16
	PS	0.09	0.24	0.39	0.701	0.05	0.07	0.23	0.31	0.757	0.04
TOQ	SF	-0.04	0.06	-0.69	0.493	-0.08	0.00	0.06	0.09	0.931	0.01
	PC	0.04	0.06	0.64	0.523	0.08	-0.04	0.06	-0.77	0.442	-0.09
	LR	0.01	0.06	0.19	0.850	0.02	-0.04	0.06	-0.76	0.449	-0.09
	AC	0.03	0.06	0.56	0.577	0.07	0.00	0.06	-0.04	0.969	0.00
	IN	0.01	0.06	0.20	0.844	0.02	0.01	0.06	0.18	0.854	0.02
	PS	0.02	0.06	0.29	0.771	0.03	-0.01	0.06	-0.16	0.875	-0.02
EF	SF	-0.82	0.93	-0.88	0.383	-0.10	-1.23	0.92	-1.35	0.183	-0.16
	PC	0.11	0.94	0.11	0.912	0.01	0.12	0.92	0.12	0.901	0.01
	LR	0.07	0.98	0.07	0.944	0.01	-0.92	0.97	-0.95	0.345	-0.11

	AC	0.04	1.01	0.04	0.966	0.01	-0.55	0.96	-0.57	0.571	-0.07
	IN	0.73	1.00	0.73	0.466	0.09	0.82	0.96	0.86	0.391	0.10
	PS	1.59	1.00	1.58	0.118	0.19	1.75	0.94	1.86	0.067	0.22
CBCL	SF	-0.06	0.09	-0.74	0.467	-0.12	-0.06	0.08	-0.83	0.413	-0.14
	PC	0.05	0.10	0.47	0.640	0.08	0.03	0.10	0.31	0.759	0.05
	LR	0.11	0.10	1.07	0.291	0.18	0.01	0.10	0.12	0.902	0.02
	AC	0.07	0.10	0.71	0.485	0.12	0.08	0.09	0.92	0.363	0.16
	IN	0.04	0.11	0.38	0.706	0.06	0.02	0.09	0.25	0.800	0.04
	PS	0.07	0.09	0.82	0.415	0.14	0.10	0.09	1.12	0.270	0.19

Note. SF=space &furnishing; PC=personal care; LR=language reasoning; AC=activity; IN=instruction; PS=program structure

Table 13

Using ECERS-R Subscale to Predict PPVT

	Alternative						Stop					
Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	-0.23	0.11	313	-2.00	.046		-0.24	0.11	313	-2.13	.034	
Fall_PPVT	0.60	0.04	313	15.34	<.001		0.59	0.04	313	15.18	<.001	
SF	0.11	0.04	69	2.55	.013	0.30	0.09	0.05	69	1.94	.057	0.23
Male	0.14	0.07	313	1.97	.050		0.14	0.07	313	2.00	.046	
ELL	0.25	0.11	313	2.27	.024		0.26	0.11	313	2.39	.017	
Subsidy	-0.30	0.08	313	-3.60	<.001		-0.29	0.08	313	-3.49	.001	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.24	69.00	819.38	851.07	-401.69		0.25	69.00	821.94	853.62	-402.97	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	-0.26	0.11	313	-2.27	.024		-0.26	0.11	313	-2.26	.025	
Fall_PPVT	0.59	0.04	313	15.04	<.001		0.59	0.04	313	15.03	<.001	
PC	0.01	0.05	69	0.29	.771	0.03	0.01	0.05	69	0.11	.911	0.01
Male	0.14	0.07	313	2.00	.046		0.14	0.07	313	2.00	.047	
ELL	0.27	0.11	313	2.49	.013		0.27	0.11	313	2.48	.014	
Subsidy	-0.27	0.08	313	-3.25	.001		-0.27	0.08	313	-3.23	.001	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.27	69.00	825.51	857.19	-404.75		0.27	69.00	825.58	857.27	-404.79	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	-0.26	0.11	313	-2.23	.026		-0.26	0.11	313	-2.24	.026	
Fall_PPVT	0.59	0.04	313	15.05	<.001		0.59	0.04	313	15.05	<.001	

LR	0.04	0.05	69	0.80	.428	0.09	0.02	0.05	69	0.52	.608	0.06
Male	0.14	0.07	313	2.00	.047		0.14	0.07	313	2.00	.046	
ELL	0.27	0.11	313	2.46	.015		0.27	0.11	313	2.46	.014	
Subsidy	-0.28	0.08	313	-3.31	.001		-0.27	0.08	313	-3.28	.001	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.27	69.00	824.95	856.64	-404.48		0.27	69.00	825.32	857.01	-404.66	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	-0.26	0.11	313	-2.25	.025		-0.26	0.11	313	-2.26	.025	
Fall_PPVT	0.59	0.04	313	15.04	<.001		0.59	0.04	313	15.04	<.001	
AC	0.01	0.05	69	0.20	.840	0.02	-0.01	0.05	69	-0.14	.893	-0.02
Male	0.14	0.07	313	1.99	.047		0.14	0.07	313	2.00	.047	
ELL	0.27	0.11	313	2.47	.014		0.27	0.11	313	2.48	.014	
Subsidy	-0.27	0.08	313	-3.24	.001		-0.27	0.08	313	-3.24	.001	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.27	69.00	825.55	857.24	-404.78		0.27	69.00	825.57	857.26	-404.79	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	-0.26	0.11	313	-2.23	.026		-0.26	0.11	313	-2.25	.025	
Fall_PPVT	0.59	0.04	313	15.02	<.001		0.59	0.04	313	15.00	<.001	
IN	0.04	0.05	69	0.81	.423	0.10	0.07	0.05	69	1.41	.163	0.17
Male	0.14	0.07	313	2.02	.045		0.14	0.07	313	2.05	.041	
ELL	0.26	0.11	313	2.43	.016		0.26	0.11	313	2.41	.017	
Subsidy	-0.27	0.08	313	-3.28	.001		-0.27	0.08	313	-3.24	.001	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.27	69.00	824.94	856.62	-404.47		0.27	69.00	823.59	855.28	-403.80	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**

Intercept	-0.26	0.11	313	-2.25	.025		-0.26	0.11	313	-2.25	.025	
Fall_PPVT	0.59	0.04	313	14.91	<.001		0.59	0.04	313	14.83	<.001	
PS	0.01	0.05	69	0.26	.795	0.03	0.01	0.05	69	0.25	.807	0.03
Male	0.14	0.07	313	2.00	.046		0.14	0.07	313	2.01	.046	
ELL	0.27	0.11	313	2.46	.015		0.27	0.11	313	2.46	.015	
Subsidy	-0.27	0.08	313	-3.23	.001		-0.27	0.08	313	-3.23	.001	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.27	69.00	825.52	857.21	-404.76		0.27	69.00	825.53	857.22	-404.77	

Note. SF=space &furnishing; PC=personal care; LR=language reasoning; AC=activity; IN=instruction; PS=program structure

Table 14

Using ECERS-R Subscale to Predict WJ

	Alternative						Stop					
Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.06	0.09	320	0.66	.509		0.05	0.09	320	0.52	.606	
Fall_WJ	0.82	0.03	320	27.55	<.001		0.81	0.03	320	27.24	<.001	
SF	0.05	0.03	70	1.69	.096	0.20	0.03	0.03	70	0.84	.406	0.10
Male	0.05	0.06	320	0.85	.398		0.05	0.06	320	0.86	.388	
ELL	-0.06	0.08	320	-0.67	.503		-0.05	0.08	320	-0.55	.581	
Subsidy	-0.11	0.07	320	-1.61	.107		-0.10	0.07	320	-1.46	.146	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.14	70	674.10	705.95	-329.05		0.15	70	676.18	708.03	-330.09	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.05	0.09	320	0.53	.595		0.05	0.09	320	0.59	.558	
Fall_WJ	0.82	0.03	320	27.55	<.001		0.82	0.03	320	27.72	<.001	
PC	-0.04	0.03	70	-1.17	.246	-0.14	-0.06	0.03	70	-1.92	.059	-0.23
Male	0.05	0.06	320	0.81	.416		0.05	0.06	320	0.84	.402	
ELL	-0.05	0.08	320	-0.58	.561		-0.05	0.08	320	-0.59	.554	
Subsidy	-0.09	0.07	320	-1.34	.181		-0.10	0.06	320	-1.47	.142	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.15	70	675.51	707.36	-329.76		0.14	70	673.23	705.09	-328.62	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.04	0.09	320	0.48	.628		0.04	0.09	320	0.50	.618	
Fall_WJ	0.82	0.03	320	27.39	<.001		0.82	0.03	320	27.46	<.001	
LR	0.02	0.03	70	0.45	.652	0.05	0.02	0.03	70	0.61	.545	0.07

Male	0.05	0.06	320	0.85	.395		0.05	0.06	320	0.85	.396	
ELL	-0.04	0.08	320	-0.53	.594		-0.05	0.08	320	-0.55	.585	
Subsidy	-0.09	0.07	320	-1.40	.162		-0.09	0.07	320	-1.42	.156	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.15	70	676.67	708.53	-330.34		0.15	70	676.51	708.36	-330.25	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.04	0.09	320	0.46	.648		0.04	0.09	320	0.49	.626	
Fall_WJ	0.82	0.03	320	27.41	<.001		0.82	0.03	320	27.46	<.001	
AC	-0.01	0.03	70	-0.32	.750	-0.04	-0.03	0.03	70	-0.77	.446	-0.09
Male	0.05	0.06	320	0.85	.396		0.05	0.06	320	0.84	.399	
ELL	-0.04	0.08	320	-0.50	.618		-0.04	0.08	320	-0.51	.614	
Subsidy	-0.09	0.07	320	-1.35	.179		-0.09	0.07	320	-1.40	.163	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.15	70	676.78	708.63	-330.39		0.15	70	676.29	708.14	-330.14	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.04	0.09	320	0.51	.614		0.04	0.09	320	0.51	.613	
Fall_WJ	0.82	0.03	320	27.39	<.001		0.81	0.03	320	27.24	<.001	
IN	0.03	0.03	70	0.72	.473	0.08	0.04	0.03	70	1.10	.274	0.13
Male	0.05	0.06	320	0.87	.385		0.05	0.06	320	0.90	.370	
ELL	-0.05	0.08	320	-0.58	.565		-0.05	0.08	320	-0.61	.545	
Subsidy	-0.09	0.07	320	-1.39	.166		-0.09	0.07	320	-1.34	.181	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.15	70	676.35	708.21	-330.18		0.15	70	675.66	707.51	-329.83	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.04	0.09	320	0.47	.639		0.04	0.09	320	0.46	.649	

Fall_WJ	0.82	0.03	320	27.42	<.001		0.82	0.03	320	27.42	<.001	
PS	0.00	0.03	70	0.06	.949	0.01	-0.02	0.03	70	-0.53	.598	-0.06
Male	0.05	0.06	320	0.84	.399		0.05	0.06	320	0.83	.409	
ELL	-0.04	0.08	320	-0.52	.605		-0.04	0.08	320	-0.47	.640	
Subsidy	-0.09	0.07	320	-1.35	.177		-0.09	0.07	320	-1.40	.161	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.15	70	676.88	708.73	-330.44		0.15	70	676.60	708.45	-330.30	

Note. SF=space &furnishing; PC=personal care; LR=language reasoning; AC=activity; IN=instruction; PS=program structure

Table 15

Using ECERS-R Subscale to Predict TEAM

	Alternative						Stop					
Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.18	0.10	314	1.70	.091		0.17	0.10	314	1.67	.096	
Fall_TEAM	0.75	0.03	314	22.63	<.001		0.75	0.03	314	22.62	<.001	
SF	0.05	0.03	68	1.43	.156	0.17	0.07	0.03	68	2.11	.039	0.25
Male	-0.13	0.07	314	-1.99	.047		-0.13	0.07	314	-1.98	.049	
ELL	-0.09	0.10	314	-0.86	.391		-0.08	0.10	314	-0.82	.413	
Subsidy	-0.11	0.07	314	-1.46	.145		-0.11	0.07	314	-1.49	.138	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.00	68	775.58	807.27	-379.79		0.00	68	773.17	804.86	-378.59	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.16	0.10	314	1.54	.125		0.16	0.10	314	1.55	.122	
Fall_TEAM	0.75	0.03	314	22.41	<.001		0.75	0.03	314	22.48	<.001	
PC	0.03	0.03	68	0.86	.395	0.10	0.02	0.03	68	0.64	.525	0.08
Male	-0.13	0.07	314	-1.96	.051		-0.13	0.07	314	-1.98	.049	
ELL	-0.07	0.10	314	-0.70	.487		-0.07	0.10	314	-0.73	.468	
Subsidy	-0.10	0.07	314	-1.39	.166		-0.10	0.07	314	-1.33	.184	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.00	68	776.93	808.61	-380.46		0.00	68	777.25	808.94	-380.63	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.17	0.10	314	1.60	.110		0.17	0.10	314	1.60	.110	

Fall_TEAM	0.75	0.03	314	22.41	<.001		0.75	0.03	314	22.53	<.001	
LR	0.02	0.03	68	0.50	.620	0.06	0.01	0.03	68	0.38	.706	0.05
Male	-0.13	0.07	314	-1.96	.051		-0.13	0.07	314	-1.96	.051	
ELL	-0.08	0.10	314	-0.78	.436		-0.08	0.10	314	-0.78	.436	
Subsidy	-0.10	0.07	314	-1.39	.166		-0.10	0.07	314	-1.37	.171	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.00	68	777.42	809.10	-380.71		0.00	68	777.52	809.21	-380.76	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.16	0.10	314	1.59	.114		0.16	0.10	314	1.58	.116	
Fall_TEAM	0.75	0.03	314	22.51	<.001		0.75	0.03	314	22.54	<.001	
AC	0.00	0.04	68	-0.10	.923	-0.01	0.02	0.03	68	0.51	.613	0.06
Male	-0.13	0.07	314	-1.96	.051		-0.13	0.07	314	-1.98	.049	
ELL	-0.08	0.10	314	-0.77	.445		-0.08	0.10	314	-0.76	.448	
Subsidy	-0.10	0.07	314	-1.34	.182		-0.10	0.07	314	-1.33	.184	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.00	68	777.66	809.35	-380.83		0.00	68	777.41	809.09	-380.70	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.17	0.10	314	1.63	.105		0.17	0.10	314	1.66	.098	
Fall_TEAM	0.75	0.03	314	22.29	<.001		0.74	0.03	314	22.05	<.001	
IN	0.03	0.04	68	0.80	.426	0.10	0.06	0.03	68	1.67	.099	0.20
Male	-0.13	0.07	314	-1.94	.053		-0.13	0.07	314	-1.90	.059	
ELL	-0.08	0.10	314	-0.82	.410		-0.09	0.10	314	-0.91	.363	
Subsidy	-0.10	0.07	314	-1.39	.167		-0.10	0.07	314	-1.31	.190	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	

Schools	0.00	68	777.02	808.70	-380.51		0.00	68	774.84	806.53	-379.42	
Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.16	0.10	314	1.59	.112		0.17	0.10	314	1.62	.107	
Fall_TEAM	0.75	0.03	314	22.43	<.001		0.75	0.03	314	22.16	<.001	
PS	0.02	0.04	68	0.45	.651	0.05	0.04	0.03	68	1.26	.211	0.15
Male	-0.13	0.07	314	-1.97	.050		-0.13	0.07	314	-1.91	.058	
ELL	-0.08	0.10	314	-0.79	.432		-0.09	0.10	314	-0.86	.391	
Subsidy	-0.10	0.07	314	-1.33	.184		-0.09	0.07	314	-1.28	.201	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.00	68	777.46	809.15	-380.73		0.00	68	776.05	807.74	-380.02	

Note. SF=space &furnishing; PC=personal care; LR=language reasoning; AC=activity; IN=instruction; PS=program structure

Table 16

Using ECERS-R Subscale to Predict LENS

	Alternative						Stop					
Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	-0.13	0.18	100	-0.76	.451		-0.12	0.18	100	-0.67	.505	
Fall_LENS	0.61	0.07	100	8.98	<.001		0.61	0.07	100	8.96	<.001	
SF	-0.07	0.07	55	-0.95	.347	-0.13	-0.07	0.07	55	-0.96	.341	-0.13
Male	0.16	0.12	100	1.41	.160		0.16	0.12	100	1.39	.169	
ELL	0.14	0.17	100	0.85	.397		0.14	0.17	100	0.83	.410	
Subsidy	0.08	0.13	100	0.58	.561		0.06	0.13	100	0.47	.637	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.28	55	361.42	386.07	-172.71		0.28	55	361.40	386.05	-172.70	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	-0.13	0.18	100	-0.76	.451		-0.12	0.18	100	-0.69	.493	
Fall_LENS	0.61	0.07	100	9.04	<.001		0.61	0.07	100	8.93	<.001	
PC	0.07	0.06	55	1.12	.268	0.15	0.03	0.06	55	0.50	.619	0.07
Male	0.17	0.12	100	1.44	.154		0.16	0.12	100	1.36	.177	
ELL	0.16	0.17	100	0.97	.336		0.14	0.17	100	0.85	.397	
Subsidy	0.04	0.13	100	0.33	.741		0.06	0.13	100	0.47	.641	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.25	55	361.12	385.78	-172.56		0.27	55	362.09	386.74	-173.04	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	-0.11	0.18	100	-0.64	.525		-0.13	0.18	100	-0.73	.467	
Fall_LENS	0.61	0.07	100	9.03	<.001		0.61	0.07	100	9.13	<.001	
LR	-0.01	0.07	55	-0.10	.917	-0.01	-0.09	0.07	55	-1.33	.188	-0.18

Male	0.15	0.12	100	1.34	.185		0.15	0.12	100	1.26	.209	
ELL	0.14	0.17	100	0.83	.409		0.16	0.16	100	0.95	.346	
Subsidy	0.06	0.13	100	0.45	.653		0.08	0.13	100	0.60	.550	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.27	55	362.34	386.99	-173.17		0.26	55	360.54	385.19	-172.27	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	-0.14	0.18	100	-0.78	.437		-0.12	0.18	100	-0.67	.504	
Fall_LENS	0.61	0.07	100	8.99	<.001		0.61	0.07	100	9.02	<.001	
AC	-0.09	0.07	55	-1.27	.208	-0.17	-0.06	0.07	55	-0.85	.400	-0.11
Male	0.17	0.12	100	1.44	.152		0.16	0.12	100	1.38	.170	
ELL	0.17	0.17	100	1.01	.316		0.15	0.17	100	0.91	.365	
Subsidy	0.06	0.13	100	0.45	.653		0.05	0.13	100	0.37	.716	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.26	55	360.71	385.36	-172.36		0.27	55	361.61	386.26	-172.81	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	-0.11	0.18	100	-0.62	.539		-0.11	0.18	100	-0.63	.531	
Fall_LENS	0.61	0.07	100	9.00	<.001		0.61	0.07	100	9.02	<.001	
IN	0.04	0.07	55	0.59	.558	0.08	0.04	0.07	55	0.53	.601	0.07
Male	0.16	0.12	100	1.40	.165		0.16	0.12	100	1.38	.170	
ELL	0.13	0.17	100	0.80	.428		0.13	0.17	100	0.80	.428	
Subsidy	0.05	0.13	100	0.35	.726		0.05	0.13	100	0.42	.678	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.27	55	361.99	386.64	-173.00		0.27	55	362.06	386.72	-173.03	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	-0.10	0.17	100	-0.55	.582		-0.09	0.17	100	-0.50	.616	

Fall_LENS	0.61	0.07	100	9.09	<.001		0.60	0.07	100	9.02	<.001	
PS	0.14	0.08	55	1.85	.070	0.25	0.14	0.07	55	1.95	.057	0.26
Male	0.16	0.11	100	1.35	.179		0.16	0.11	100	1.40	.163	
ELL	0.11	0.16	100	0.66	.513		0.10	0.16	100	0.62	.536	
Subsidy	0.06	0.13	100	0.46	.649		0.05	0.13	100	0.41	.680	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.24	55	358.98	383.63	-171.49		0.24	55	358.63	383.29	-171.32	

Note. SF=space &furnishing; PC=personal care; LR=language reasoning; AC=activity; IN=instruction; PS=program structure

Table 17

Using ECERS-R Subscale to Prdict EWA-Word

	Alternative						Stop					
Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.14	0.11	323	1.28	.202		0.15	0.11	323	1.31	.190	
Fall_Word	0.66	0.04	323	17.27	<.001		0.66	0.04	323	17.31	<.001	
SF	-0.01	0.05	70	-0.19	.847	-0.02	0.02	0.04	70	0.50	.621	0.06
Male	0.16	0.07	323	2.35	.019		0.16	0.07	323	2.35	.019	
ELL	-0.26	0.11	323	-2.43	.016		-0.26	0.11	323	-2.46	.015	
Subsidy	-0.04	0.08	323	-0.49	.624		-0.05	0.08	323	-0.55	.584	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.26	70	868.30	900.21	-426.15		0.25	70	868.09	900.00	-426.05	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.15	0.11	323	1.32	.188		0.14	0.11	323	1.28	.202	
Fall_Word	0.66	0.04	323	17.33	<.001		0.66	0.04	323	17.29	<.001	
PC	-0.04	0.05	70	-0.96	.339	-0.11	0.02	0.04	70	0.49	.623	0.06
Male	0.16	0.07	323	2.33	.020		0.16	0.07	323	2.34	.020	
ELL	-0.26	0.11	323	-2.48	.014		-0.26	0.11	323	-2.43	.016	
Subsidy	-0.04	0.08	323	-0.46	.648		-0.04	0.08	323	-0.50	.617	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.26		867.40	899.31	-425.70		0.26	70	868.09	900.01	-426.05	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.15	0.11	323	1.33	.184		0.15	0.11	323	1.31	.190	
Fall_Word	0.66	0.04	323	17.33	<.001		0.66	0.04	323	17.28	<.001	
LR	0.04	0.05	70	0.80	.427	0.09	0.03	0.05	70	0.55	.584	0.06

Male	0.16	0.07	323	2.35	.019		0.16	0.07	323	2.35	.020	
ELL	-0.26	0.11	323	-2.48	.014		-0.26	0.11	323	-2.46	.014	
Subsidy	-0.05	0.08	323	-0.58	.562		-0.05	0.08	323	-0.55	.583	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.25	70	867.70	899.61	-425.85		0.25	70	868.03	899.95	-426.02	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.14	0.11	323	1.30	.194		0.14	0.11	323	1.30	.196	
Fall_Word	0.66	0.04	323	17.28	<.001		0.66	0.04	323	17.39	<.001	
AC	0.02	0.05	70	0.33	.739	0.04	0.06	0.05	70	1.30	.199	0.15
Male	0.16	0.07	323	2.34	.020		0.16	0.07	323	2.33	.020	
ELL	-0.26	0.11	323	-2.46	.015		-0.26	0.11	323	-2.49	.013	
Subsidy	-0.04	0.08	323	-0.52	.606		-0.04	0.08	323	-0.48	.629	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.26	70	868.23	900.14	-426.11		0.25	70	866.64	898.55	-425.32	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.14	0.11	323	1.29	.198		0.15	0.11	323	1.34	.182	
Fall_Word	0.66	0.04	323	17.27	<.001		0.66	0.04	323	17.25	<.001	
IN	0.00	0.05	70	0.00	.998	0.00	0.04	0.05	70	0.85	.401	0.10
Male	0.16	0.07	323	2.35	.019		0.17	0.07	323	2.38	.018	
ELL	-0.26	0.11	323	-2.43	.016		-0.27	0.11	323	-2.52	.012	
Subsidy	-0.04	0.08	323	-0.51	.612		-0.04	0.08	323	-0.49	.622	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.26	70	868.34	900.25	-426.17		0.25	70	867.62	899.54	-425.81	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.15	0.11	323	1.36	.175		0.15	0.11	323	1.40	.163	

Fall_PPVT	0.66	0.04	323	17.39	<.001		0.66	0.04	323	17.53	<.001	
PS	0.06	0.05	70	1.15	.253	0.14	0.09	0.05	70	2.07	.042	0.24
Male	0.16	0.07	323	2.33	.020		0.17	0.07	323	2.39	.017	
ELL	-0.27	0.11	323	-2.55	.011		-0.28	0.11	323	-2.64	.009	
Subsidy	-0.04	0.08	323	-0.48	.635		-0.04	0.08	323	-0.43	.666	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.25	70	867.01	898.92	-425.50		0.24	70	864.12	896.03	-424.06	

Note. SF=space &furnishing; PC=personal care; LR=language reasoning; AC=activity; IN=instruction; PS=program structure

Table 18

Using ECERS-R Subscale to Predict EWA-Name

	Alternative						Stop					
Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	3.81	0.64	319	5.94	.000		3.81	0.64	319	5.96	.000	
Fall_Name	0.64	0.05	319	13.72	<.001		0.64	0.05	319	13.74	<.001	
SF	0.28	0.22	70	1.24	.218	0.15	0.38	0.22	70	1.71	.092	0.20
Male	0.21	0.38	319	0.55	.581		0.22	0.38	319	0.58	.562	
ELL	-0.31	0.56	319	-0.56	.576		-0.30	0.56	319	-0.54	.593	
Subsidy	-0.22	0.44	319	-0.50	.615		-0.24	0.44	319	-0.55	.584	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	1.04	70	2183.62	2215.45	-1083.81		1.00	70	2182.29	2214.12	-1083.14	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	3.75	0.64	319	5.84	<.001		3.73	0.64	319	5.82	<.001	
Fall_Name	0.64	0.05	319	13.64	<.001		0.64	0.05	319	13.74	<.001	
PC	0.13	0.23	70	0.56	.575	0.07	0.32	0.22	70	1.48	.144	0.17
Male	0.22	0.38	319	0.57	.567		0.21	0.38	319	0.55	.585	
ELL	-0.26	0.56	319	-0.45	.651		-0.25	0.56	319	-0.45	.653	
Subsidy	-0.19	0.44	319	-0.42	.673		-0.16	0.44	319	-0.37	.708	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	1.08	70	2184.84	2216.67	-1084.42		1.03	70	2182.99	2214.82	-1083.50	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	3.77	0.64	319	5.86	<.001		3.77	0.64	319	5.85	<.001	
Fall_Name	0.64	0.05	319	13.65	<.001		0.64	0.05	319	13.65	<.001	
LR	0.09	0.23	70	0.37	.714	0.04	0.01	0.23	70	0.03	.976	0.00

Male	0.22	0.38	319	0.57	.572		0.21	0.38	319	0.56	.575	
ELL	-0.28	0.56	319	-0.50	.617		-0.28	0.56	319	-0.49	.626	
Subsidy	-0.19	0.44	319	-0.42	.672		-0.17	0.44	319	-0.40	.692	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	1.08	70	2185.03	2216.86	-1084.51		1.09	70	2185.16	2217.00	-1084.58	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	3.77	0.64	319	5.85	<.001		3.75	0.64	319	5.83	<.001	
Fall_Name	0.64	0.05	319	13.64	<.001		0.65	0.05	319	13.69	<.001	
AC	0.01	0.24	70	0.06	.950	0.01	0.16	0.23	70	0.71	.479	0.08
Male	0.21	0.38	319	0.56	.577		0.21	0.38	319	0.54	.587	
ELL	-0.28	0.56	319	-0.49	.625		-0.29	0.56	319	-0.51	.611	
Subsidy	-0.17	0.44	319	-0.40	.692		-0.17	0.44	319	-0.38	.707	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	1.09	70	2185.16	2216.99	-1084.58		1.07	70	2184.65	2216.48	-1084.33	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	3.78	0.64	319	5.86	<.001		3.82	0.64	319	5.95	<.001	
Fall_Name	0.64	0.05	319	13.62	<.001		0.64	0.05	319	13.61	<.001	
IN	0.06	0.24	70	0.27	.789	0.03	0.31	0.23	70	1.37	.175	0.16
Male	0.22	0.38	319	0.57	.569		0.24	0.38	319	0.63	.528	
ELL	-0.29	0.57	319	-0.51	.612		-0.34	0.56	319	-0.61	.542	
Subsidy	-0.18	0.44	319	-0.40	.687		-0.16	0.44	319	-0.36	.722	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	1.08	70	2185.09	2216.92	-1084.55		1.02	70	2183.32	2215.15	-1083.66	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	3.77	0.64	319	5.86	<.001		3.77	0.64	319	5.86	<.001	

Fall_Name	0.64	0.05	319	13.66	<.001		0.64	0.05	319	13.65	<.001	
PS	0.09	0.24	70	0.39	.701	0.05	0.07	0.23	70	0.31	.757	0.04
Male	0.21	0.38	319	0.55	.579		0.22	0.38	319	0.57	.569	
ELL	-0.29	0.57	319	-0.52	.604		-0.29	0.57	319	-0.51	.610	
Subsidy	-0.16	0.44	319	-0.37	.709		-0.16	0.44	319	-0.37	.711	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	1.09	70	2185.01	2216.85	-1084.51		1.09	70	2185.07	2216.90	-1084.53	

Note. SF=space &furnishing; PC=personal care; LR=language reasoning; AC=activity; IN=instruction; PS=program structure

Table 19

Using ECERS-R Subscale to Predict TOQ

	Alternative						Stop					
Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.14	0.15	301	0.90	.370		0.15	0.15	301	0.96	.336	
Fall_TOQ	0.20	0.05	301	3.73	<.001		0.20	0.05	301	3.67	<.001	
SF	-0.04	0.06	70	-0.69	.493	-0.08	0.00	0.06	70	0.09	.931	0.01
Male	0.09	0.10	301	0.88	.380		0.09	0.10	301	0.88	.380	
ELL	-0.16	0.15	301	-1.06	.289		-0.16	0.15	301	-1.13	.262	
Subsidy	-0.19	0.12	301	-1.67	.096		-0.20	0.12	301	-1.74	.084	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.24	70	1061.63	1093.08	-522.81		0.24	70	1062.10	1093.56	-523.05	
Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.14	0.15	301	0.95	.344		0.15	0.15	301	0.98	.330	
Fall_TOQ	0.20	0.05	301	3.70	<.001		0.20	0.05	301	3.69	<.001	
PC	0.04	0.06	70	0.64	.523	0.08	-0.04	0.06	70	-0.77	.442	-0.09
Male	0.09	0.10	301	0.88	.379		0.09	0.10	301	0.89	.374	
ELL	-0.16	0.15	301	-1.10	.273		-0.17	0.15	301	-1.13	.258	
Subsidy	-0.20	0.11	301	-1.77	.078		-0.20	0.12	301	-1.75	.081	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.23	70	1061.70	1093.16	-522.85		0.25	70	1061.51	1092.96	-522.75	
Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.15	0.15	301	0.97	.333		0.14	0.15	301	0.92	.360	
Fall_TOQ	0.20	0.05	301	3.68	<.001		0.20	0.05	301	3.72	<.001	
LR	0.01	0.06	70	0.19	.850	0.02	-0.04	0.06	70	-0.76	.449	-0.09

Male	0.09	0.10	301	0.88	.382		0.09	0.10	301	0.89	.376	
ELL	-0.17	0.15	301	-1.13	.259		-0.16	0.15	301	-1.08	.281	
Subsidy	-0.20	0.12	301	-1.74	.082		-0.19	0.12	301	-1.67	.096	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.24	70	1062.07	1093.53	-523.04		0.24	70	1061.52	1092.98	-522.76	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.15	0.15	301	0.99	.322		0.15	0.15	301	0.96	.338	
Fall_TOQ	0.20	0.05	301	3.69	<.001		0.20	0.05	301	3.69	<.001	
AC	0.03	0.06	70	0.56	.577	0.07	0.00	0.06	70	-0.04	.969	0.00
Male	0.09	0.10	301	0.85	.396		0.09	0.10	301	0.88	.380	
ELL	-0.17	0.15	301	-1.15	.250		-0.16	0.15	301	-1.12	.264	
Subsidy	-0.20	0.11	301	-1.76	.080		-0.20	0.12	301	-1.74	.084	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.24	70	1061.79	1093.25	-522.89		0.24	70	1062.11	1093.56	-523.05	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.15	0.15	301	0.97	.331		0.15	0.15	301	0.97	.333	
Fall_TOQ	0.20	0.05	301	3.66	<.001		0.20	0.05	301	3.65	<.001	
IN	0.01	0.06	70	0.20	.844	0.02	0.01	0.06	70	0.18	.854	0.02
Male	0.09	0.10	301	0.88	.379		0.09	0.10	301	0.88	.377	
ELL	-0.17	0.15	301	-1.14	.256		-0.17	0.15	301	-1.14	.257	
Subsidy	-0.20	0.12	301	-1.74	.083		-0.20	0.12	301	-1.73	.085	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.24	70	1062.07	1093.53	-523.03		0.24	70	1062.07	1093.53	-523.04	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.15	0.15	301	0.98	.330		0.14	0.15	301	0.95	.342	

Fall_TOQ	0.20	0.05	301	3.70	<.001		0.20	0.05	301	3.69	<.001	
PS	0.02	0.06	70	0.29	.771	0.03	-0.01	0.06	70	-0.16	.875	-0.02
Male	0.09	0.10	301	0.87	.385		0.09	0.10	301	0.87	.383	
ELL	-0.17	0.15	301	-1.15	.253		-0.16	0.15	301	-1.10	.272	
Subsidy	-0.20	0.12	301	-1.72	.086		-0.20	0.12	301	-1.74	.083	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.24	70	1062.02	1093.48	-523.01		0.24	70	1062.08	1093.54	-523.04	

Note. SF=space &furnishing; PC=personal care; LR=language reasoning; AC=activity; IN=instruction; PS=program structure

Table 20

Using ECERS-R Subscale to Predict EF

	Alternative						Stop					
Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	12.26	2.50	306	4.90	<.001		12.28	2.49	306	4.93	<.001	
Fall_EF	0.60	0.05	306	11.76	<.001		0.60	0.05	306	11.83	<.001	
SF	-0.82	0.93	68	-0.88	.383	-0.10	-1.23	0.92	68	-1.35	.183	-0.16
Male	0.00	1.50	306	0.00	.999		-0.02	1.50	306	-0.01	.991	
ELL	1.53	2.33	306	0.66	.513		1.49	2.33	306	0.64	.522	
Subsidy	-3.40	1.78	306	-1.91	.057		-3.38	1.77	306	-1.91	.058	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	4.88	68	3137.53	3169.06	-1560.77		4.80	68	3136.49	3168.01	-1560.24	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	12.32	2.51	306	4.91	<.001		12.31	2.51	306	4.91	<.001	
Fall_EF	0.60	0.05	306	11.83	<.001		0.60	0.05	306	11.82	<.001	
PC	0.11	0.94	68	0.11	.912	0.01	0.12	0.92	68	0.12	.901	0.01
Male	0.00	1.50	306	0.00	1.000		0.00	1.50	306	0.00	.998	
ELL	1.46	2.34	306	0.62	.534		1.45	2.33	306	0.62	.535	
Subsidy	-3.53	1.78	306	-1.99	.048		-3.52	1.78	306	-1.98	.048	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	4.88	68	3138.30	3169.82	-1561.15		4.88	68	3138.30	3169.82	-1561.15	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	12.33	2.50	306	4.93	<.001		12.36	2.50	306	4.95	<.001	
Fall_EF	0.60	0.05	306	11.80	<.001		0.60	0.05	306	11.71	<.001	
LR	0.07	0.98	68	0.07	.944	0.01	-0.92	0.97	68	-0.95	.345	-0.11

Male	0.00	1.50	306	0.00	.999		-0.01	1.50	306	-0.01	.995	
ELL	1.44	2.33	306	0.62	.538		1.48	2.33	306	0.64	.525	
Subsidy	-3.53	1.78	306	-1.98	.048		-3.42	1.78	306	-1.93	.055	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	4.90	68	3138.31	3169.83	-1561.16		4.84	68	3137.40	3168.92	-1560.70	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	12.33	2.50	306	4.93	<.001		12.42	2.51	306	4.96	<.001	
Fall_EF	0.60	0.05	306	11.75	<.001		0.60	0.05	306	11.67	<.001	
AC	0.04	1.01	68	0.04	.966	0.01	-0.55	0.96	68	-0.57	.571	-0.07
Male	0.00	1.50	306	0.00	.998		0.01	1.50	306	0.01	.995	
ELL	1.44	2.33	306	0.62	.538		1.44	2.33	306	0.62	.538	
Subsidy	-3.52	1.78	306	-1.98	.048		-3.56	1.78	306	-2.00	.046	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	4.89	68	3138.31	3169.83	-1561.16		4.89	68	3137.99	3169.51	-1560.99	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	12.33	2.50	306	4.94	<.001		12.30	2.50	306	4.93	<.001	
Fall_EF	0.61	0.05	306	11.88	<.001		0.60	0.05	306	11.88	<.001	
IN	0.73	1.00	68	0.73	.466	0.09	0.82	0.96	68	0.86	.391	0.10
Male	0.02	1.50	306	0.01	.989		0.05	1.50	306	0.03	.975	
ELL	1.37	2.33	306	0.59	.559		1.36	2.33	306	0.58	.560	
Subsidy	-3.57	1.77	306	-2.02	.045		-3.47	1.77	306	-1.96	.051	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	4.81	68	3137.78	3169.30	-1560.89		4.78	68	3137.57	3169.09	-1560.79	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	12.35	2.49	306	4.97	<.001		12.44	2.48	306	5.02	<.001	

Fall_EF	0.61	0.05	306	11.96	<.001		0.61	0.05	306	11.95	<.001	
PS	1.59	1.00	68	1.58	.118	0.19	1.75	0.94	68	1.86	.067	0.22
Male	-0.06	1.50	306	-0.04	.971		0.04	1.50	306	0.03	.980	
ELL	1.25	2.32	306	0.54	.591		1.14	2.32	306	0.49	.625	
Subsidy	-3.45	1.76	306	-1.95	.052		-3.38	1.76	306	-1.92	.056	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	4.60	68	3135.84	3167.36	-1559.92		4.51	68	3134.93	3166.45	-1559.47	

Note. SF=space &furnishing; PC=personal care; LR=language reasoning; AC=activity; IN=instruction; PS=program structure

Table 21

Using ECERS-R Subscale to Predict CBCL

	Alternative						Stop					
Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.22	0.17	165	1.31	.193		0.22	0.17	165	1.28	.202	
Fall_CBCL	0.75	0.05	165	13.65	<.001		0.75	0.05	165	13.63	<.001	
SF	-0.06	0.09	35	-0.74	.467	-0.12	-0.06	0.08	35	-0.83	.413	-0.14
Male	-0.10	0.09	165	-1.11	.268		-0.10	0.09	165	-1.12	.264	
ELL	-0.24	0.15	165	-1.56	.120		-0.24	0.15	165	-1.58	.116	
Subsidy	0.14	0.12	165	1.21	.228		0.15	0.12	165	1.24	.217	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.48	35	430.94	457.57	-207.47		0.48	35	430.80	457.42	-207.40	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.21	0.17	165	1.25	.213		0.22	0.17	165	1.29	.199	
Fall_CBCL	0.75	0.06	165	13.68	<.001		0.75	0.05	165	13.68	<.001	
PC	0.05	0.10	35	0.47	.640	0.08	0.03	0.10	35	0.31	.759	0.05
Male	-0.10	0.09	165	-1.14	.256		-0.10	0.09	165	-1.14	.257	
ELL	-0.24	0.15	165	-1.58	.116		-0.25	0.16	165	-1.60	.111	
Subsidy	0.13	0.12	165	1.15	.252		0.14	0.12	165	1.18	.241	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.48	35	431.27	457.89	-207.63		0.48	35	431.39	458.02	-207.70	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.21	0.17	165	1.23	.222		0.21	0.17	165	1.26	.208	
Fall_CBCL	0.75	0.05	165	13.71	<.001		0.75	0.05	165	13.67	<.001	

LR	0.11	0.10	35	1.07	.291	0.18	0.01	0.10	35	0.12	.902	0.02
Male	-0.10	0.09	165	-1.15	.252		-0.10	0.09	165	-1.14	.256	
ELL	-0.25	0.15	165	-1.62	.108		-0.25	0.15	165	-1.59	.115	
Subsidy	0.13	0.12	165	1.14	.257		0.14	0.12	165	1.17	.245	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.48	35	430.32	456.94	-207.16		0.49	35	431.48	458.10	-207.74	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.21	0.17	165	1.26	.211		0.21	0.17	165	1.24	.215	
Fall_CBCL	0.75	0.05	165	13.66	<.001		0.75	0.06	165	13.59	<.001	
AC	0.07	0.10	35	0.71	.485	0.12	0.08	0.09	35	0.92	.363	0.16
Male	-0.10	0.09	165	-1.17	.245		-0.10	0.09	165	-1.14	.254	
ELL	-0.25	0.15	165	-1.60	.112		-0.25	0.15	165	-1.61	.109	
Subsidy	0.14	0.12	165	1.19	.236		0.14	0.12	165	1.19	.237	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.48	35	430.98	457.60	-207.49		0.48	35	430.62	457.25	-207.31	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**
Intercept	0.21	0.17	165	1.26	.210		0.22	0.17	165	1.27	.205	
Fall_CBCL	0.75	0.05	165	13.69	<.001		0.75	0.06	165	13.67	<.001	
IN	0.04	0.11	35	0.38	.706	0.06	0.02	0.09	35	0.25	.800	0.04
Male	-0.10	0.09	165	-1.14	.258		-0.10	0.09	165	-1.14	.257	
ELL	-0.25	0.16	165	-1.61	.110		-0.25	0.16	165	-1.60	.112	
Subsidy	0.14	0.12	165	1.18	.240		0.14	0.12	165	1.18	.240	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.48	35	431.34	457.97	-207.67		0.48	35	431.43	458.05	-207.71	

Fixed Effect	*Coeff*	*SE*	*df*	*t*	*p*	*d**	*Coeff*	*SE*	*df*	*t*	*p*	*d**

Intercept	0.22	0.17	165	1.31	.194		0.22	0.17	165	1.32	.190	
Fall_CBCL	0.75	0.05	165	13.70	<.001		0.75	0.05	165	13.72	<.001	
PS	0.07	0.09	35	0.82	.415	0.14	0.10	0.09	35	1.12	.270	0.19
Male	-0.10	0.09	165	-1.14	.257		-0.10	0.09	165	-1.12	.263	
ELL	-0.25	0.15	165	-1.63	.105		-0.26	0.15	165	-1.67	.098	
Subsidy	0.14	0.12	165	1.17	.244		0.14	0.12	165	1.17	.244	
Random Effect	*Var*	*df*	*AIC*	*BIC*	*Log*		*Var*	*df*	*AIC*	*BIC*	*Log*	
Schools	0.48	35	430.80	457.42	-207.40		0.48	35	430.22	456.84	-207.11	

Note. SF=space &furnishing; PC=personal care; LR=language reasoning; AC=activity; IN=instruction; PS=program structure

Table 22

Theme by Coaches

Coach	Theme 1	Theme 2	Theme 3
Amy	X	X	X
Anne	X		X
Benny	X	X	X
Carta	X	X	X
Daisy	X	X	X
Janet	X	X	
Jane			X
Katy	X	X	X
Nancy	X	X	X
Rayann	X	X	X
Tara	X	X	X
Vicky		X	X
Yan	X	X	X

Appendix A Coach Interview Protocol

The focus of this interview is to gather your feedback about using the environmental rating scale (ERS) to support your coaching work.

1. I am very interested about your work. Could you briefly introduce your current coaching roles and responsibilities? How is that look like in your daily work?

2. Could you also describe about the centers, the child care providers (i.e., program directors or teachers), and the children's characteristics that you work with?

3. Which version (s) of ERS do you use in your coaching? How do you use the tool in your coaching work?

4. I learned that you use the "stop score" and "scoring all the way up" for different purposes. Can you describe more about that?

5. Could you give me an example about the process of goal setting using ERS? What are the steps you would normally take?

6. Have you encounter any issues related to using the ERS to set goals? What kind of support or resources you may turn to? What additional support you would want?

7. Since you have received the reliability training of the ERS tool, how do you feel about the training? Any suggestions for improvement?

Appendix B Qualitative Code

1. Tool-Childcare Program Provider Dimension

1.1 Item feasibility: the quote mentions the indicators' feasibility and constrain when implementing in the real-world programs.

1.2 Culture adaptation: the quote mentions whether the items are adaptable and appropriate with different cultural practices and program philosophies.

1.3 Score interpretation by providers: the quote describes how program providers make sense of the score results.

1.4 Convergent validity: the quote compares ERS scores with other assessment tools (e.g., CLASS) and program quality guidance.

1.5 Assessment procedure: the quote describes the assessment procedure and its implication for program provider.

2. Tool-Coach Dimension

2.1 Personal belief: the quote captures coaches' personal belief about the assessment.

2.2 Score interpretation by coaches: the quote describes how coach make sense of the score.

2.3 Scoring preference: the quote describes coaches' preference in using the "stop" and "all the way up" scoring rules.

2.4 Coaching needs: the quote mentions professional development and support needed to better use the tool.

2.5 Agency: the coach describes her own confidence and capacity to use the tool.

3. Coach-Childcare Program Provider Dimension

3.1 Strength-based coaching: the quote describes coach's belief about using a strength-based approach to coaching.

3.2 Partnership building: the quotes describes the importance of partnership building before using other practice-based coaching practices (e.g., goal setting, feedback).

3.3 Goal setting: the quotes describes goal setting process and data-based decision making practices.

4. Others: quotes that do not fit any of the criteria above