Classroom Quality in the Age of Accountability: Using a Comprehensive Multidimensional Rasch Approach to Investigate the Validity of the Early Childhood Environment Rating Scale-Revised

A Dissertation submitted by

Brandon Foster

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

in Child Study and Human Development

Tufts University

Eliot-Pearson Department of Child Study and Human Development

May 2017

Committee: Drs. Christine McWayne (Chair),

Tama Leventhal, Shelagh Peoples, and Eric Dearing

Abstract

The Early Childhood Environment Rating Scale–Revised (ECERS-R) is the most widely used measure of classroom quality, and has been implemented into numerous states' Quality Rating and Improvement Systems (QRISs). However, the ECERS-R was never designed for implementation into these systems or to be used for the purposes of accountability. As such, its validity for determining benchmarks on key indicators of classroom quality and suitability for identifying low or high performing preschools has yet to be established. Findings from a handful of validity studies which utilized classical test theory (CTT) have highlighted information pertaining to the factor structure of the measure. However, the CTT approach provides very limited diagnostic information about the functioning of the ECERS-R (e.g., Fan, 1998; Kieffer, 1998; Macdonald & Paunonen, 2002; Traub, 1997; Schumacker & Smith, 2007), and studies investigating the psychometric properties of the ECERS-R using more rigorous methods from item-response theory have hardly emerged (Gordon et al., 2013; 2015). This dissertation provides the first comprehensive account of the psychometric properties of the ECERS-R, utilizing the ECLS-B dataset, a nationally representative sample of early childhood environments. In particular, three proposed dimensional specifications of the measure that have emerged from the literature were examined. Psychometric analyses were carried out using Andrich Rating Scale Models and Multidimensional Random Coefficient Multinomial Logit (MRCML) models. To organize this information, this dissertation adopted Wolfe and Smith's (2007) Rasch validity framework, which ensured that analyses provided psychometric information about each component of Messick's seminal (1989) construct validity framework. Results illuminate the psychometric trade-offs that occur when different specifications of the ECERS-R are utilized to measure classroom quality. Additional information, which builds off the work of Gordon et al. (2013; 2015b), is provided to guide policymakers in the use of total scores from the measure in policy applications. Further, a promising new dimensional structure of the ECERS-R is posited and confirmed. Results from these analyses suggested that the ECERS-R functioned as a risk assessment, capable of precisely measuring classrooms with low levels of classroom quality, but functioned poorly for classrooms with average or above average levels of classroom quality. These results raise serious concerns about the use of the ECERS-R in most states' early-childhood assessment systems, and call into question some basic assumptions for how this measure is assumed to have performed in the prior research studies.

Acknowledgements

The culmination of this milestone would not have been possible without the support of many individuals.

I would like to first thank my dissertation committee. Drs. Christine McWayne, Tama Leventhal, and Shelagh Peoples have each provided helpful feedback and have been great mentors.

To my advisor, Dr. Christine McWayne, for her unwavering support and encouragement over the years. She has helped me realize professional potentials that I could have never imagined, and always kept me focused on the big picture. Her passion for this work continues to inspire me.

To my EP Ph.D. cohort (Cricket, Lisette, Maggie and Judith), for their friendship and laughter. I was not expecting to make such amazing friends while here, and I am so grateful for the support they have provided me.

To labmates (Sunah & Lok), for their kindness and generosity.

To my sister, Kayla Foster, for her love and support, which always came just in the nick of time. She has never let me forget to smile throughout this process, as well as in life.

To my parents, Theresa and Ron Foster, who answered every call and assuaged every doubt. It was their hard work and continued sacrifices when I was growing up that enabled me to reach beyond what I thought was possible. This accomplishment is as much theirs as it is mine. I know they are proud, and that fills my heart with joy.

And finally, to my wonderful wife, Julie Thomason, who has been there with me since the beginning of this journey. My gratitude for her sacrifices is unmeasurable. Her honesty and tough love helped me to dig deeper, work harder, and persist. I love you and could not imagine a better partner in life.

Thank you.

Table of Contents

List of Tables

List of Figures

Introduction

Policymakers have increasingly focused on investments in education as a way for laying the foundation of America's 21st century economy. Societal factors like the increasing cost of childcare, increased maternal labor force participation (Sall, 2014), and a focus on educational equity for all children (Burkam, 2013), have coalesced with concerns about the preparedness of U.S. children to meet the demands of a fluid and technology-based economic landscape (Education & Workforce, 2008). In an effort to address these societal concerns, policymakers have moved to provide unrestricted access to preschool for every child, based most recently on evidence by Heckman (2006; 2010), which has highlighted the strong return on investment in preschool programming. Indeed, the role of high quality preschool classrooms has become a linchpin among solutions to improve our country's school readiness efforts (Yoshikawa, Weiland, Brooks-Gunn, Burchinal, Espinosa, Gormley & Zaslow, 2013).

This emphasis on early childhood education (ECE) settings is pragmatic, as ECE is theorized to be one of the most important proximal influences on children's development (Bronfrenbrenner & Morris, 1998; Weisner, 2002). It is theorized that high-quality ECE settings offer a multitude of learning opportunities that have the potential to promote children's early educational successes in positive ways. These kinds of opportunities include, but are not limited to: exposure to positive peer interactions (NIHCD, 2001), early vocabulary, reading and mathematics concepts, rich instructional materials, and nurturing teachers who facilitate positive adult and peer interactions to optimize learning (Mashburn, Pianta, Hamre, Downer, Barbarin, Bryant & Howes, 2008). The empirical literature has continually highlighted small, positive direct effects for preschool exposure on children's later educational outcomes (Gormley, Gayer,

Phillips, & Dawson, 2005; Hustedt, Barnett, Jung, & Goetze, 2009; Hustedt, Barnett, Jung, & Thomas, 2007; Weiland & Yoshikawa, 2013; Wong, Cook, Barnett, & Jung, 2008).

Despite research supporting the impact of preschool on children's development, the statistics surrounding preschool utilization in the U.S. show a broad unmet need. First, the U.S. lags behind many developed countries in the world in the rate with which children 3 to 4 years of age are enrolled in formal educational programs (i.e., preschool). The Organization for Economic Cooperation and Development (OECD), in their annual report titled *Education at a Glance* (2015), show that, across 31 developed countries, the average enrollment rate of 3- to 4-year olds in formal schooling is 81%. Among the list of 31 developed countries, the United States ranked in the bottom three in terms of the number of children enrolled in preschool, with only 59% of 4-year olds and 41% of 3-year olds enrolled in preschool. In order to mitigate these crucial gaps, the U.S. is currently investing 6.2 billion dollars into preschool programs. That figure is up by 10% ($573 million) from the 2013-2014 year (NIEER The State of Preschool Yearbook, 2015).

Increasingly, as tax payer dollars are being used to expand access to preschool, accountability efforts have been commensurate (Schultz, 2015). Most of these accountability efforts have focused on ensuring that preschools are meeting expectations for quality. For example, in an effort to increase current levels of quality, some states have set up incentives for programs to engage in continuous improvement efforts, typically by tying reimbursement for services to providers' scores for quality (Pianta, 2012). The success of these efforts is contingent on the ability to measure classroom quality in a way that is both accurate and comprehensive. As such, many states have co-opted extant observational measures of classroom quality into their accountability efforts – with the most utilized measure being the Early Childhood Environment Rating Scale-Revised (ECERS-R) (National Center on Childcare Quality Improvement, 2013).

Policymakers have assumed measures like the ECERS-R are valid for these purposes, and this provides some assurance that the data gathered are relevant to outcomes that are the impetus for investments into preschool. However, measures like the ECERS-R were not designed for this purpose, and scholars are increasingly questioning the validity of observational measures like the ECERS-R (Goldstein & Flake, 2016; Gordon, 2013; 2015b; Votruba-Drzal & Miller, 2016).

Currently, there is a dearth of comprehensive and rigorous psychometric evidence to support the assumptions underlying the use of measures like the ECERS-R in accountability efforts (Charalambous, Blazar, McGinn, Kraft, Beisiegel, Humez & Lynch, 2012; Gordon, 2013; 2015b; Pianta, 2012). Further, studies that have investigated the validity of the ECERS-R have failed to situate the work within larger validity frameworks. As such, findings from research have led to equivocal conclusions, and psychometric information about the measure is difficult for policymakers to access and utilize. Situating psychometric analyses of the ECERS-R in larger validity frameworks provides a systematic way to document a range of evidence which can be used to support the use of the ECERS-R for policymaking. The goal of this dissertation is address these deficits. As such, a primary aim of these analyses is to provide both researchers and policymakers with comprehensive information about the psychometric properties of the ECERS-R that can be used to either support or disprove a range of claims about the validity of the ECERS-R for its varied uses in policy applications (Charalambous et al., 2012).

The forthcoming sections will first situate the work in theoretical conceptualizations for classroom quality. An overview of relevant literature will then be summarized. This will build to critical gaps in the literature, which are followed by the research questions for the study. This is followed by a description of the methodology used to address research questions, where Wolfe & Smith's (2007) Rasch validity framework is used to situate and organize analyses in line with

Messick's (1989) construct validity framework. Next, results are presented. In the final section conclusions and policy implications are discussed, which is followed by limitations and future directions for research.

Guiding Theoretical Frameworks for Understanding Preschool Classroom Quality

Measures of classroom quality can be described theoretically using the bioecological model (Bronfrenbrenner & Morris, 1998). Within this theory, children's learning and development, teacher pedagogy and practices, and classroom environments can be examined as interrelated levels which are part of a larger developmental system. Each level of the developmental system has the potential to impact children's learning and development through varying degrees of proximity to a child's lived-in experience. Specifically, this theory provides a framework for understanding how microsystems (e.g., family and school contexts), mesosystems (e.g., the interactions between various microsystems), exosystems (e.g., neighborhood and school districts, etc.) and macrosystems (e.g., economic and social policies of a nation) can impact children's development and learning.

At the macrosystem level, local, state, and federal policies can influence preschool quality, and subsequently impact children's development. For example, education policy is a reflection of a society's ideologies and value systems (Ball, 2006). This can be seen in the educational reform efforts in the U.S. that have occurred over the last two decades, which have favored expanding access to high-quality preschool for all children in the U.S. The aggregate momentum of these efforts has been in the service of increasing the school readiness skills of children in the U.S. A natural consequence of these efforts at the national level, is that certain pedagogical environments and practices have become favored.

The exosystem level entails the environments that affect children's everyday experience – those systems that indirectly influence children, but are nonetheless distal from their day-to-day interactions. At this level, children's families can play a critical role in shaping their development, primarily through their decisions about the appropriateness of a given childcare

arrangement for their children. Meyers and Jordan (2006) use an accommodations framework to describe how variations in the quality of care children are exposed to is driven by parental decisions about childcare. This framework highlights the importance of contextualized decisions about childcare using three dimensions: parents' a priori preferences and tastes for quality, social networks as a source for information, and parents' perceptions of available supply and resources for obtaining care. As parents navigate decisions about childcare, they construct their beliefs and preferences for childcare through accommodating trade-offs between the conditions necessary for their continued participation in the labor market, and optimal care for their children. For most parents this is not a simple consumption choice, as childcare is an infrequent purchase in the lives of most people. Further, information about the quality of childcare options is often lacking, causing the cost incurred by parents when making decisions about childcare to come with considerable risk. As a consequence, parents turn to their social networks for signals to guide their decision-making. These social networks are valuable, because they function as heuristics that are filled with cultural information and values, which aids in limiting the range of acceptable/normative options for care that parents consider. However, this limited pool of acceptable childcare options is also impacted by the supply of childcare. Parents do not decide among similar options, as factors like family income, proximity of diverse center types, and hours of operation, etc. can limit the pool of available childcare options. In sum, parents navigate these contexts in making consumption choices about childcare arrangements, and the choices parents make about childcare can lead to variation in the quality of care children receive.

Primary caregivers (i.e., teachers) are also part of a child's exosystem. The characteristics of caregivers also indirectly determine the quality of preschool environments that children are exposed to on a regular basis. For example, childcare centers with more support for teachers'

professional development might moderate the quality of their interactions with children in the classroom (Pianta, DeCoster, Cabell, Burchinal, Hamre, Downer, LoCasale-Crouch, Wilford & Howes, 2014; Yoshikawa, Leyva, Snow, Treviño, Barata, Weiland, Gomez, Moreno, Rolla, D'Sha, & Arbour, 2015). Further, administrative and other supports a center provides to a teacher might also indirectly influence classroom quality (Goelman & Guo, 1998; Wells, 2017). Other teacher characteristics, such as teacher education level, have shown associations with classroom quality (Burchinal, Cryer, Clifford & Howes, 2002; Howes, Whitebook & Phillips, 1992; NICHD ECCRN, 2002; Scarr, Eisenberg & Deater-Deckard, 1994). Further, teacher perception variables, such as their sense of self-efficacy, might also indirectly influence the quality of care children receive (Guo, Piasta, Justice & Kaderavek, 2010).

The microsystem entails contexts with the moment-to-moment interactions between teachers and children. At this level, interactions between children and teachers becomes a proximal driver of children's development of school readiness skills. In fact, teacher-child interactions are hypothesized to be the "primary mechanism by which children learn in the classroom" (cf. in Curby, Rimm- Kaufman, & Cameron Ponitz, 2009, p. 913). For example, complex interactions between children and educational materials in preschool environments likely mediate their interaction with teachers (Kontos, Burchinal, Howes, Wisseh & Galinsky, 2002; Kontos & Keyes, 1999). Further, research has demonstrated that levels of positive interactions between children and preschool teachers are associated with higher levels of motivation and engagement to learn in children (Howes, Burchinal, Pianta, Bryant, Early, Clifford, Barbarin, 2008; Ladd, Birch, & Buhs, 2003; Pianta, Steinberg, & Rollins, 1995). Interactions between children and ECE teachers have also been shown to be dynamic, unfolding in complex ways over time, with children's engagement associated with later teacher emotional

and organizational supports (Curby, Downer & Booren, 2014). This is a finding which indicates that children might also aide in helping shape the quality of the preschool environments they are exposed to through their active participation and engagement with teachers, who in turn reshape the environment to respond to the needs of the children in their classrooms.

Constructionist learning theories are also useful for understanding the importance of the type and quality of early childhood environments. Socio-cultural theories, like Vygotsky's (1979), contend that mediated interactions between a child, and either a teacher or a more competent peer, in the use of materials for learning, can help a child move to more advanced levels of cognitive functioning. Within Vygotsky's theoretical framework, high quality classrooms contain caregivers who are sensitive in their interactions with children. Teachers create a high-quality classroom through their scaffolded interactions with children, whereby teachers are attentive to the individual needs of children in their classroom, and structure their interactions with the explicit purpose of supporting children in developing increasingly complex knowledge and skills. A critical tool that teachers utilize in these interactions is complex language, which encourages children to reason through their experiences in a classroom. Caregivers in these classrooms are sensitive to the need for continuous monitoring of the current skill-level of children in their classrooms so that these interactions can be modified. Furthermore, constructivist theories of learning (Piaget, 1952) stress the importance of children's active exploration and manipulation of their environments. As part of this active exploration and manipulation, children play the role of "little scientists" who form cognitive schemes to understand their lived-in experiences, and refine these through testing new hypotheses. Using constructivist theories as a framework for understanding high quality preschool classrooms, classrooms filled with open-ended activities and manipulatives that are organized in such a way

as to encourage autonomy and exploration are optimal (Harms, Clifford, & Cryer, 2005; Stipek & Byler, 2004).

In sum, notions of what constitutes preschool classroom quality are influenced by several theories of child development. Constructionist theories have coalesced to emphasize several core features of high quality classrooms. Specifically, these theories have stressed the sensitivity and warmth of caregivers, the availability of developmentally appropriate materials and activities in the classroom, the time allotted for unstructured engagement with those materials and environments, and the importance of sensitive caregivers who scaffold their interactions with children to support children moving through increasingly complex levels of cognitive development (Burchinal, Magnuson, Powell & Hong, 2015). Further, systems theories, like Bronfrenbrenner & Morris's (1998) bioecological model and Meyers and Jordan's (2006) accommodations framework, assert the important role of the interaction between various contextual factors in shaping the quality of preschool environments to which children are exposed.

Literature Review

Current Policy Context for Expanding Access to High Quality Early Childhood Education in the U.S.

The importance of early childhood education within the landscape of education policy efforts in the United States has expanded considerably under the reauthorization of the Elementary and Secondary Education Act (ESEA), which was first signed into law by President Johnson in 1965. The ESEA was reauthorized and amended with the Every Student Succeeds Act (ESSA), which was signed into law by President Obama in 2015. ESSA shifted the discussion of education policy away from K-12 to a focus on P-12, and provided states, districts, and schools with several important policy levers to both expand access and improve the quality of early childhood education programs.

Primarily, ESSA provided states, LEAs, and SEAs with considerable flexibility in using Title I funds to support the learning and needs of children before they enter kindergarten, thereby creating a powerful lever that could be used to strategically expand access to preschool. For example, Title I, Part A, gave schools the ability to use all, or a portion, of their Title I funds to provide preschool for eligible students (i.e., low income students). Under ESSA SEAs are also allowed to expand the pool of students who have access to free or subsidized preschool, by allowing Title I funds to be used for any student in a school if at least 40% of students in that school are from low-income families. An LEA may reserve a portion of funds from its Title I allocation to operate a preschool program for eligible children in the LEA as a whole or in a portion of the LEA. Further, LEAs are authorized to use Title 1 funds to improve the quality or extend the number of days children spend in childcare programs. SEAs and LEAs were also given considerable flexibility in using Title I funds to improve the quality of care, by providing early childhood teachers with access to high quality professional development. Finally, Title IV,

Part X of the ESEA allowed Title I money to be allocated to charter schools that provide access to preschool, provided these schools also provide either elementary and/or secondary education.

Further, in an effort to expand the quality of early childhood education programs, ESSA stressed the importance of alignment, collaboration, and coordination between early childhood education programs and the K-12 system. The law encouraged states and LEAs to be thoughtful about the consistency and connectedness of both programs and professional standards across contexts that serve the same grade-level of children (i.e., horizontal alignment). For example, the State plans that are required under ESEA for the use of Title I funds require states to coordinate efforts across programs providing preschool to children, which includes programs administered under other departments like Head Start. The Department of Education issued non-regulatory guidance which urged states to think about how the quality of early childhood programming could be improved through aligning different early childhood programs within a State. Further, states were also encouraged to think about how their early childhood programs vertically aligned with state K-12 systems. As a policy aim, ESSA encouraged states to adopt a strategic P–3 approach to early childhood programing. Through this approach, states were encouraged to either document or develop a framework for what children should know prior to entering kindergarten. Taken in sum, both means of alignment are designed to encourage activities that are likely to increase the quality of programs (U.S. Department of Education, 2016).

Ahead of ESSA becoming law, congress authorized funds to catalyze and incentivize the process of vertically and horizontally aligning early childhood programs within states. These funds were made available for two competitive grant programs, specifically the Early Learning Challenge program (ELC), which was part of the Race to the Top (R2T) challenge (Race to the Top Act, 2011), and the Preschool Development and Expansion grants, which were authorized in

2014. Both of these grant programs influenced the alignment activities recommended to states by the Department of Education as part of their non-regulatory guidance for implementing ESSA (U.S. Department of Education 2016). Previously, the R2T ELC required states to design efforts to implement integrated systems to ensure that their state's preschool programming was of high quality. Forty states applied for these grants, and 20 received awards. States were urged to adopt cross-sector Quality Rating and Improvement Systems (QRIS) to both, rate the quality of providers, and track children's progress as they moved from preschool into the K-12 pipeline. Further, states were encouraged to link children's progress to their teachers in order to build an evidence base for what was and was not effective in early childhood classrooms. The Preschool Development and Expansion grants that were authorized ahead of ESSA becoming law were designed specifically to expand access to high quality preschool for low-to moderate-income students. Thirty states applied for these grants, and 18 were awarded. Building off of the R2T ELC, the grant process required the awardees to have or develop a QRIS. However, the Preschool Development and Expansion grants went further, in that qualified providers who received support from these grants were required to participate in their state's QRIS.

The QRIS National Learning Network, an organization that helps states implement QRISs, defines a QRIS as "an intentionally transparent definition of the progression of program quality from basic to excellent" (Schilder, Iruka, Dichter, & Mathias, 2015). A QRIS is a systemic approach to assess, improve, and communicate the level of quality of early-childhood programs. QRISs are composed of five common elements: (1) Program standards that are used to assign ratings to participating providers; (2) Supports for programs and practitioners, typically in the form of technical assistance, to help programs in their continuous improvement efforts; (3) Financial incentives for providers to participate in the QRIS and engage in continuous

improvement efforts; (4) Quality accountability and monitoring processes that are used to set benchmarks to determine how providers are meeting or exceeding expectations; (5) Consumer education that is used to provide stakeholders with key information to aid in the decision-making process about which programs are good fits for the needs of their children. A logic model that outlines the basic functioning of QRISs can be found in Figure 1.

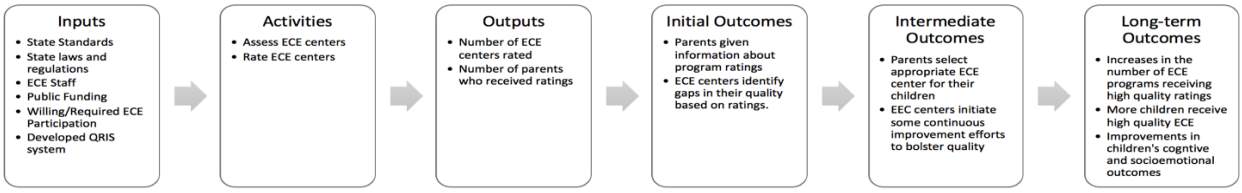

Figure 1. *Sample logic model for QRIS functioning, which is adapted from McCawley (2001) and the W. K. Kellogg Foundation (2004)*

The premise underlying QRISs is to provide parents with the information they need about the quality of early childhood programs, so they can "vote with their feet" as to which program best fits the needs of their child. The goal of providing power to parents in the form of specific information about quality of programs is that it should incentivize programs to engage in continuous improvement efforts to bolster their program quality, and as a result expand the number of high quality programs available to families within states. (Buettner & Andrews, 2009). As a result of the R2T ELC and the Preschool Development and Expansion grants, every state in America has or is creating a QRIS (Tout, Starr, Soli, Moodie, Kirby & Boller, 2010).

Both the R2T and the Preschool Development and Expansion grants gave states considerable control over defining their standards for quality used in QRISs, but required that states create systems for tracking quality across time. As states rushed to implement QRIS systems, policymakers raced to find effective indicators for program quality. As a result, many states rushed to adopt extant measures of early childhood classroom quality into their QRISs (Tout et al., 2010). The most commonly implemented measures have been the Classroom

Assessment Scoring System (CLASS) and the Early Childhood Environment and Rating Scale-Revised (ECERS-R; NIEER The State of Preschool Yearbook, 2015), both of which are observational measures, with the CLASS emphasizing more of the teacher-interactional components of quality, and the ECERS-R emphasizing more the structural aspects of program quality. To date, the ECERS-R remains the most widely implemented measure of classroom environment implemented into these systems (NIEER The State of Preschool Yearbook, 2015), with 30 states adopting the ECERS-R into their programs of assessment and evaluation (Child Trends, 2014; Mashburn et al., 2008).

As the utilization of QRISs has increased rapidly, scholars have called into question whether QRISs function as intended. As a consequence, a complex picture of the efficacy of these systems has emerged. Specifically, links between measures of quality embedded in QRISs and child outcomes have been mixed. For example, Sabol and Pianta (2015) found no relations between QRIS measures of quality and child academic outcomes. Another study found no evidence between QRIS measures of quality and children's socioemotional outcomes (Hestenes et al., 2015). Increasingly, researchers are calling attention to the need to validate QRISs by examining the validity of components of QRISs, not limited to a thorough vetting of the psychometrics pertaining to ratings and associated measures of quality embedded within QRISs. Many of these measures of ECE quality were never designed to be used for the purpose of program accountability (Lahti, Elicker, Zellman & Fiene, 2014). Given the widespread use of the ECERS-R in QRISs, it is critical to ensure that rigorous and comprehensive information about the psychometric properties of this measure is available to stakeholders and policymakers.

Summary of the Context for Investigating the Validity of Measures of Preschool Quality

Public policy pertaining to early childhood programing over the last decade has increasingly emphasized the role of high quality preschool in relation to child outcomes. High quality early childhood environments have been a central focus in these efforts, in part because of the well documented positive associations between these environments and child outcomes (Gormley, et al., 2005; Hustedt, et al., 2007; Hustedt, et al., 2009; Weiland & Yoshikawa, 2013; Wong, Cook, Barnett, & Jung, 2008). As such, policymakers are looking to leverage indicators of ECE quality to positively impact children's school readiness— school-entry cognitive (i.e., mathematics, language and reading), attentional, and socioemotional skills (Duncan, Dowsett, Claessens, Magnuson, Huston, Klebanov & Sexton, 2007). The passing of ESSA into law in 2015 marked a critical point in these policy efforts, as the law gave states levers to expand access to preschool to a larger number of children using Federal dollars. Further, many states are utilizing QRISs to provide parents with information about the quality of preschool options, and also using these as a tool to improve the quality of early childhood programming within states. However, scholars are increasingly highlighting potential issues with the use of QRISs for both formative and summative purposes (Hestenes et al., 2015; Sabol & Pianta, 2015; Le, et al., 2015). In doing so, an urgent need has emerged to investigate the validity of measures of quality embedded in these QRIS systems, as many of the existing instruments designed to measure quality in early childhood classrooms were not initially developed to be used for accountability purposes (Lahti, et al., 2014). Given that nearly every state in the U.S. is now implementing a QRIS system, and that most of these systems are utilizing extant ECE measures for quality, it is important to establish validity evidence for the myriad purposes for which measures like the ECERS-R are currently being used in the U.S.'s educational policy efforts.

Definitions of Early Childhood Classroom Quality

As the United States moves towards expanding access to high quality preschool to all children, the field has faced a considerable challenge in defining preschool quality (La Paro, Thomason, Lower, Kitner-Duffy, Cassidy, 2012). This is in part because the definitions of quality are a matter of perspective (Katz, 1993; Layzer & Goodson, 2006). For example, parents tend to conceptualize quality as an accommodations tradeoff between environments they feel will meet the needs of their children, and the flexibility of programs to accommodate their individual work schedules (Cryer and Burchinal, 1997; Emlen, 1999). Additionally, teachers' perspectives about quality tend to focus more on their working conditions, with the idea that improved working conditions likely set the stage for better instruction, and better outcomes in children (Phillips, Howes & Whitebook, 1991). However, the child's perspective has dominated most conceptualizations of how classroom quality is defined and measured (Layzer & Goodson, 2006). Within this perspective, definitions of quality concern features of the classroom, interactions with teachers and peers, and contextual factors, that are likely to indirectly impact teachers' interactions with children, and subsequently promote positive outcomes in children. As such, conceptualizations of classroom quality have typically stressed the importance of both structural and process aspects of quality.

Structural quality was originally conceptualized by Phillips and Howes (1987) to identify features of a classroom that could be regulated by policy efforts, and were likely to promote positive child outcomes. Over the course of a decade, scholars built on the work of Phillips and Howes (1987) to define several important structural characteristics associated with quality and outcomes. Specifically, scholars posited the importance of characteristics like staff qualifications and training (Arnett, 1989, Peisner-Feinberg, Burchinal, Clifford, Culkin, Howes, Kagan,

Yazejian, Byler, Rustici & Zelazo, 2000; Burchinal et al., 2002; de Kruif, McWilliam, Ridley & Wakely, 2000; Howes, Whitebook, & Phillips, 1992; NICHD ECCRN, 2002; Scarr, Eisenberg & Deater-Deckard, 1994), stability of caregivers (Howes and Hamilton, 1992; Whitebook & Sakai, 2003), and adult-to-child ratios (NICHD Early Childcare Research Network, 1996; Kontos, Howes, Shinn, and Galinsky, 1995).

Cryer (1999) expanded the work of Phillips and Howes (1987) by drawing on Bronfrenbrenner's ecological systems theory (Bronfenbrenner, 1992). Cryer and colleagues (1999) depicted process quality (i.e., the day-to-day interactional episodes between children and teachers) at the center of ECE care, surrounded by structural quality indicators. Cryer and colleagues (1999) expanded the scope of structural indicators for quality by highlighting the importance of considering the presence of materials and activities for learning, ease of access for materials for learning, and personal care routines as sources of classroom quality. However, Cryer and colleagues (1999) envisioned these structural indicators for quality as indirectly influencing the more proximal day-to-day interactions in a classroom that were likely to promote children's development. Cassidy and colleagues (2012) further advanced the conceptualization of structural indicators of quality by defining them as "independent of human interaction between individuals" (pg. 511). In doing so, the researchers aligned their work with Cryer et al. (1999), by affirming the notion that structural indicators could be defined along a continuum of proximal relatedness to a child's everyday situated experience in the classroom, highlighting the importance of indicators for structural quality like materials for learning, activities and the utilization of space in a classroom. However, the researchers went further by making it clear that the purpose of defining structural indicators for ECE quality was to document the existence of features in the classroom that had the potential to catalyze rich dynamic interactions between

children and teachers, which could in turn promote children's development. This conceptualization of structural quality suggests that the presence of structural indicators of quality in a classroom likely represents a minimal condition for program quality, but can help define a high-quality classroom if, and only if, they catalyze and aid in sustaining dynamic interactions between children and teachers in the classroom.

Definitions of process quality have also undergone changes since they were first introduced by Philips and Howes (1987) who defined it as "[the] dynamic environment that captures children's actual experiences in childcare" (pg. 9). Vandell and Wolfe (2000) described process quality further as features that "combine experiences across several areas that include health and safety provisions, interactions with caregivers, and age-appropriate materials" (p. 3). As a result, indicators like materials for learning and physical space in the classroom could be considered process indicators; however, as argued above, these same indicators could also be considered proximal indicators for structural quality. Again, the work of Cassidy and colleagues (2005) has helped to move to the field to a clearer differentiation of process and structural quality. In their qualitative analysis of classroom quality literature, the researchers posited that process quality indicators "require human interaction among individuals" (pg. 510). In this view, the teacher or other children become the vehicles for the process quality indicators, which are mediated by things like the physical space in the room or materials for learning. As such, indicators like teacher language use, behavior modeling, and instructional episodes can become indicators for ECE process quality.

Hamre and Pianta (2007), building on the work of Philips and Howes (1987), and Dunn (1993), have also provided a framework for understanding process quality. The researchers stressed the importance of interactions between teachers and children as the primary mechanism

through which children develop school readiness skills. As a result, within this framework, indicators for process quality are exclusive to the interactional episodes between teachers and children. Their framework organized interactions between teachers and children under three main constructs– Emotional Support, Classroom Organization, and Instructional Support. Teacher Emotional Support is thought to positively support children's development through teacher's successful facilitation of positive teacher-student and student-student interactions (Hamre & Pianta, 2001; Harter, 1996; Ladd, Birch, & Buhs, 1999; Pianta, Steinberg, & Rollins, 1995; Roeser, Eccles, & Sameroff, 2000; Ryan, Stiller, & Lynch, 1994; Silver, Measelle, Essex, & Armstrong, 2005; Wentzel, 1998). The Classroom Organization construct is influenced by research showing the importance of managing student conduct, behavior, time and attention in classrooms (Blair, 2002; Connell & Wellborn, 1991; Grolnick & Ryan, 1989; Rimm-Kaufman, Early, Cox, Saluja, Pianta, Bradley & Payne, 2002; Rubin, Coplan, Fox, & Calkins, 1995; Tobin & Graziano, 2006). The Instructional Support dimension draws on the notion that teachers provide children with opportunities to demonstrate existing skills, and scaffold children to develop increasingly complex skills (Davis & Miyake, 2004; Vygotsky, 1991).

In sum, conceptualizations of preschool classroom quality have evolved to reflect both structural and process aspects of quality. The understanding of structural components of preschool classroom quality have largely been thought of as features of the preschool environment that could be regulated by policymakers (Philips & Howes, 1987). This view has evolved to encompass the aspects of classroom environments that are independent of human interaction between individuals (Cryer et al., 2012). These aspects of classroom quality can entail teacher-level factors, such as their education and training, as well as program-level features that might impact a teacher and their subsequent instruction. These variables might include the

availability of professional development or other provisions for staff needs. Further, variables that represent space and furnishings in the classroom or the presence of activities and materials for learning might also be considered structural aspects of classroom quality. The conceptualizations of process aspects of classroom quality have come to represent the interactions among children and teachers (Cassidy et al., 2005; Hamre & Pianta, 2007). The theory underlying this component of classroom quality is that these interactions are both the basis by which children evidence their current level of development to teachers, and the tools through which teachers scaffold their interactions with children to move them through increasingly complex levels of development. As researchers have come to define preschool classroom quality with more precision, they have become increasingly interested in measuring these features of the classroom.

Measuring Classroom Quality

As conceptualizations of classroom quality crystalized in the literature, scholars increasingly focused on designing ways to measure structural and process quality features in the classroom. Layzer & Goodson (2007) describe three families of measurement for classroom quality: measures of structural care, measures of process care, and global measures that combine both structural and process features of care.

Traditional measures of structural quality have come in the form of surveys that are administered to gather information about centers and classrooms (Layzer & Goodson, 2007), and are typically found in all national datasets pertaining to early childhood care. These surveys are designed for both teachers and parents. Often times in these surveys teachers are asked to self-report on key professional and center characteristics, such as highest level of education, experience with courses in early childhood, compensation, etc. Parents are often asked to report

information pertaining to the amount of time children spend in childcare, satisfaction with the arrangement, etc. However, these surveys are often viewed as inferior to more established measures for quality due to lack of documented validity evidence.

Measures of process quality focus on the interactions between children and teachers and/or children and other children in the classroom. All of these measures were designed to capture the essence of a child's experience in the classroom across contexts. This is because early childhood environments are theorized to be among the most important, proximal influences on children's development, learning, and education (Bronfenbrenner & Morris, 1998; Weisner, 2002). High quality care settings are thought to be filled with learning opportunities that have the potential to promote children's early educational successes. These measures are usually administered by independent observers. Several examples of these measures are found in the literature, such as: the Arnett Caregiver Interaction Scale (Arnett, 1989), Observational Record of the Caregiving Environment (NICHD ECCRN, 1996 & 2001), and the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2007). Scholars believe that these measures are the most reflective of the proximal classroom processes likely to relate to child outcomes (McCabe & Ackerman, 2007).

Measures of global classroom quality are also well established in the literature for classroom quality. These measures have been designed to measure both structural and process quality in early childhood classrooms. As such, the measures take a "whole-child" perspective to understanding classroom quality, within which materials and activities for learning, personal care routines, and space for learning are critical for understanding classroom quality, but so are the pedagogical interactions that involve these activities and spaces. Like measures of process quality, these measures are typically administered by trained and independent observers. Several

domains of classroom quality are established by items, which are rated by observers on a variant of likert scale. Several examples of these measures exist in the literature, such as: The Early Childhood Classroom Observation Measure (Spek & Byler, 2004), Assessment Profile for Early Childhood Programs (Abbott-Shim & Sibley, 1998), and the Early Childhood Environment Rating Scale-Revised (Harms, Clifford & Cryer, 1998). The goal of these measures is to capture the *average* experience of children in a classroom. Measures of global quality are among the most utilized measures of early childhood classroom quality.

Relationships Between Measures of ECE Quality and Child Outcomes

Measures of classroom quality have been a critical tool for understanding important policy questions about early childhood care and access. Consequently, researchers have focused on understanding how these measures predict child outcomes, and whether certain features of the early childhood classroom are more likely specific outcomes in children.

Two measures of classroom quality have held sway in studies seeking to answer these questions, the ECERS-R and CLASS. Both have been widely employed in several national datasets pertaining to ECE quality and care (e.g., the National Center for Early Development and Learning's [NCEDL]; Multi-State Study of Pre-Kindergarten [Multi-State Study]; the NCEDL – NIEER State-Wide Early Education Programs Study [SWEEP Study]; and the Preschool Curriculum Evaluation Research [PCER] Initiative Study). Both measures were designed to represent a different aspect of ECE quality, with the ECERS and its revised version, the Early Childhood Environment Rating Scale-Revised (ECERS-R) (Harms et al., 1998) designed to be a global measure of classroom quality that taps into several structural and process indicators indicative of overall ECE quality, and the CLASS designed to tap into the interactions between teachers and children. In particular, three meta-analyses have been influential for understanding

associations between these measures and child outcomes (Burchinal, Kainz, Cai, Tout, Zaslow, Martinez-Beck & Rathgeb, 2009; Burchinal, Kainz, & Cai, 2011; Keys, Farkas, Burchinal, Duncan, Vandell, Li, Ruzek & Howes, 2013).

A set of studies by Burchinal and colleagues (2009; 2011) utilized a meta-analysis (i.e., N = 20 studies) to summarize the effect of observational measures of classroom quality on child outcomes. After the authors' inclusion criteria, 97 associations between measures of classroom quality and child outcomes were meta-analyzed. Results yielded a modest range of partial correlations (i.e., $r = 0.05$ to $r = 0.17$) between the observational measures (i.e., the ECERS-R and CLASS) and child outcomes, with the strongest associations between these measures and children's cognitive outcomes. Another meta-analysis by Keys, Farkas, Burchinal, Duncan, Vandell, Li, Ruzek and Howes (2013) found similar results using a set of four prominent national datasets and examining school readiness skills (i.e., children's language, mathematics and social skills) at kindergarten entry. The authors found that relations between measures of classroom quality, such as the ECERS-R and the CLASS, and child outcomes were between about $r = 0.03$ to $r = 0.05$ in magnitude. Again, the strongest effects were concentrated in children's cognitive skills, with no relations between these measures and children's socio-emotional outcomes observed.

A study by Mashburn et al. (2008) used data from the National Center for Early Development and Learning's (NCEDL) Multi-State Study of Pre-Kindergarten (Multi-State Study) and the NCEDL –NIEER State-Wide Early Education Programs Study (SWEEP Study found that the total score for the ECERS-R showed small positive relations with children's expressive language scores, while the instructional support factor from the CLASS showed small positive associations with children's expressive, receptive and applied problems scores,

respectively. Additionally, the emotional support factor from the CLASS predicted increases in children's social competence and decreases in their problem behaviors. A similar study by Howes and colleagues (2008), utilized the same datasets, but considered the ECERS-R scores across two factors – a Teaching Provisions factor (pertaining to the language use and interactional episodes between children and teachers in the classroom), and a Provisions for Learning factor that measured the presence of materials for learning, and the use of space in the classroom. The researchers found associations between the ECERS-R Teaching and Interactions Scale and both expressive (d = .06) and receptive language (d = .08). Additionally, the Instructional Climate factor from the CLASS was associated with identifying letters (d = .07), and math skills (d = .06), while the Emotional Climate factor was associated with increases in math skills (d = .05). The results suggest that the instructional support factor from the ECERS-R and the Instructional Climate factor of the CLASS were associated with the largest number of child outcomes, most of which were linguistic in nature.

The results using these measures in other national datasets have been more mixed. For example, Auger, Farkas, Burchinal, Duncan and Vendell (2014) used the Preschool Curriculum Evaluation Research (PCER) Initiative Study with the ECERS-R and an instrumental variable design to show that the teaching and interactions factor of the ECERS-R was positively associated to a small degree with children's math scores, while the provisions for learning factor was associated with both math and language scores. Similarly, Gordon, Fujimoto, Kaestner, Korenman & Abner, (2013) used the ECERS-R with the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) dataset and found no relations between ECERS-R total scores or factor scores and children's cognitive and socioemotional outcomes.

In sum, investigations into the relationship between measures of classroom quality and child outcomes have primarily utilized the ECERS-R and the CLASS. Studies using the CLASS and the ECERS-R have utilized scores from the measures to explore these relationships in different ways. Researchers have examined associations between the CLASS and child outcomes using a multidimensional approach with both an Instructional and Emotional Climate factor. The ECERS-R has been used as both a global measure of classroom quality with the total score from the measure, but multidimensional approaches have also been considered using an Instructional Support and Provisions for Learning factor. The most consistent relationships between these measures and child outcomes are for the portions of both measures that tap into the curriculum driven interactional episodes between children and teachers (i.e., the Instructional Support factor from the ECERS-R and the Instructional Climate factor from the CLASS), both of which relate to measures of children's cognitive skills. However, associations between these measures and child outcomes have been small— a result that has been replicated not just across individual studies, but also using meta-analyses. This is surprising, as developmental theory suggests that these measures, specifically the interactional components of these measures, should predict child outcomes. Yet findings from research would suggest that these specific measures are likely not suitable for understanding child outcomes. Despite this, these measures have been widely co-opted into disparate policy applications. As a result, scholars are increasingly advocating for more thorough investigations into the psychometric properties of measures like the ECERS-R (Gordon, Fujimoto, Kaestner, Korenman & Abner, 2013; Gordon, Hofer, Fujimoto, Risk, Kaestner & Korenman, 2015b; Layzer & Goodson, 2006; Mashburn, 2017; Perlman, Zellman, & Le, 2004).

History of the ECERS

The original version of the ECERS was developed by Harms & Clifford (1980) as a global measure of classroom quality. The activities involved in designing the content for the original version of the ECERS utilized an empirical foundation to ensure items were developed in line with best practices from early childhood research (Harms & Clifford, 1980; Harms et al., 1998). As such, the development of the measure emphasized a whole-child perspective, which called for an integration of cognitive, emotional and behavioral aspects of children's development. Further, the items from the measure emphasized the importance of child-initiated activities that are guided by carefully trained education professionals. The measure utilized a checklist of 37 items, which were grouped together to reflect seven unique aspects of classroom quality. The ECERS was designed to be completed by a trained observer (Sakai, Whitebook, Wishard & Howes, 2003). The total score for each item was established on a 7-point scale (1 = "inadequate" to 7 = "excellent"), which reflected increasing levels of quality for each item. Items from the original ECERS were designed to align with the following subscales: Personal Care, Furnishings and Display for Children, Language-Reasoning and Experiences, Fine and Gross Motor Activities, Creative Activities, Social Development, and Adult Needs.

In 1998, the ECERS underwent a revision and became the ECERS-R (Harms et al., 1998). Revisions were made, in line with the National Association for the Education of Young Children's (NAEYC) guidelines for developmentally appropriate practice. The guidelines highlighted the importance of safety of care, opportunities for children to direct their own engagement in a variety of activities, and instructional support to promote children's learning and development within these activities, as well as positive interactions with teachers (Bredekamp & Copple, 1997). Developers of the ECERS-R revised the measure through: (1)

Examining the content of the ECERS in relation to other measures for early childhood classroom environments, (2) Examining extant research studies utilizing the ECERS, and (3) eliciting feedback from stakeholders who had utilized the measure. Consequently, the ECERS was expanded from 37 to 43 items, which were still organized within seven distinct subscales that were designed to reflect: Space and Furnishings, Personal Care Routines, Language-Reasoning, Activities, Interaction, Program Structure, and Parents and Staff. Significant revisions to the measure included the elimination of redundant items and addition of more relevant items for the underlying subscales. This process was accomplished through collaboration between practitioners and researchers in small focus groups. In addition, subscales were renamed to better align the content of the items (Sakai, Whitebook, Wishard & Howes, 2004). Descriptions of the items aligning to each subscale of both the ECERS and ECERS-R are found in Table 1.

Table 1. *Comparison of items on both forms of the ECERS*

ECERS (1982)	ECERS-R (1998)
Personal care	*Personal routines*
• Greeting/departing • Meals/snacks • Nap/rest • Diapering/toileting • Personal grooming	• Greeting/departing • Meals/snacks • Nap/rest • Toileting/diapering • Health practices • Safety practices
Furnishings and display for children	*Space and furnishings*
• For routing care • For learning activities • For relaxation and comfort • Room arrangement • Child-related display	• Indoor space • Furniture care for play and learning • Furnishings for relaxation • Room arrangement • Space for privacy • Child-related display • Space for gross motor • Gross motor equipment
Language-reasoning and experiences	*Language-reasoning*
• Understand of language • Using language • Using learning concepts • Informal use of language	• Books and pictures • Encouraging children to communicate • Reasoning skills • Informal use of language
Fine and gross motor activities	*Activities*
• Perceptual/fine motor	• Fine motor

• Supervision of fine motor activities • Space for gross motor • Gross motor equipment • Time for gross motor activities • Supervision of gross motor activities	• Art • Music/movement • Blocks • Sand/water • Dramatic play • Nature/science • Math/number • Use of TV, video and/or computers • Promoting acceptance of diversity
Creative activities	*Program structure*
• Art • Music/movement • Blocks • Sand/water • Dramatic play • Schedule • Supervision of creative activities	• Schedule • Free play • Group time • Provisions for children with disabilities
Social development	*Interaction*
• Space to be alone • Free play • Group time • Cultural activities • Tone • Provisions for exceptional children	• Supervision of gross motor activities • General supervision of children • Discipline • Staff-child interactions • Interactions among children
Adult Needs	*Parents and Staff*
• Adult personal area • Opportunities for professional growth • Adult meeting area • Provisions for parents	• Provisions for parents • Provisions for personal staff needs • Provisions for personal needs of staff • Staff interaction and cooperation • Supervision and evaluation of staff • Opportunities for professional growth

Notes. From Sakai, Whitebook, Wishard & Howes, 2004.

Investigations Into the Validity of the ECERS-R

Given the widespread adoption of the ECERS-R in policy applications, interest has been paid to understanding the validity of the measure. To date, most studies have investigated the structural validity of the measure. The goal of these analyses has been to understand whether the ECERS-R strays from its intended design as a global measure of classroom quality, and instead represents several underlying dimensions of quality. In order to explore issues concerning the multidimensionality of the measure, most researchers have relied on factor analytic techniques. In particular, three studies have been cited widely as key sources of evidence to support claims about validity of the ECERS-R.

Using principal components analysis and a sample size of 68 classrooms, Sakai and colleagues (2004) explored the dimensional structure of both the ECERS and ECERS-R. The goal of their work was to explore whether the dimensional structure of the measure was stable across both versions of the measure. Across both of the ECERS the researchers found evidence for a two-factor solution. The content of the items that aligned to each of the factors reflected Teaching Interactions (i.e., items pertaining to interactions between teachers/staff and children) and Provisions for Learning (i.e., items pertaining to the presence of materials and activities for learning in a classroom). The factor solution for both of these factors fit well with theoretical conceptualizations of quality, as each contained content that was aligned with distinct aspects of structural and process quality, respectively. Factor loadings for the Teaching and Interactions factor ranged from .49 to .87, and exhibited an alpha of .84, while the Provisions for Learning factor contained items with factor loadings ranging from .51 to .82, and exhibited an alpha of .81. Further, the results suggested that a shorter version of the ECERS-R was sufficient, as only 19 of the 43 items contained factor loadings that were of a magnitude that warranted items being retained in the final factor solutions. Additionally, the researchers conceptualized that each of these factors could be used as unique measures of quality, and as such tested the discriminant validity and found that item-to-total correlations were higher for items with their respective factors than for the other factor (e.g., items for the Provisions for Learning factor correlate higher for the total score of that factor than for the Teaching and Interactions factor). In its own right, this result that should not have been surprising, as the explicit purpose of factor analysis is to maximize inter-item correlations within factors. It would have been more useful for Sakai and colleagues (2004) to examine how different the item-to-total correlations looked for each item within their respective factor when compared to the item-to-total correlations of the measure

when used as a global indicator for classroom quality. However, it raises an important issue that was not explored in the analyses of Sakai and colleagues (2004), namely how different would a classroom look on the continuum for quality if the total score for the measure was used as the metric for quality versus the dimensional scores for each factor? Said in other terms, how much would the multidimensional representation of the measure represent substantive differences in underlying levels of classroom quality when compared to the total score for the measure? Finally, the researchers also examined the convergent validity of the total score from the ECERS-R by examining associations with the Arnett Caregiver Interactions Scale (Arnett, 1989). In doing so, the researchers found associations in the expected direction, with teacher sensitivity positively associated with total scores from the ECERS ($r = .60$) and ECERS-R ($r = .54$), and teacher harshness was negatively associated with total scores from the ECERS ($r = -.56$) and ECERS-R ($r = -.52$).

That same year, Perlman, Zellman and Le (2004) also used principal components analysis to examine the dimensional structure of the ECERS-R in a larger sample of classrooms ($N = 326$). The researchers observed similar patterns for the dimensional structure of the measure as Sakai and colleagues (2004), but yet reached opposite conclusions about the multidimensionality of the measure. Initial exploratory analyses showed items from the measure reflected three factors. Similar to Sakai et al (2004), the first factor contained items for child activities, program structure, and space and furnishings. The second factor contained items measuring staff–child interactions, including personal care routines and the encouragement of language development. A third factor containing items pertaining to provisions for parents and staff. Despite this, Perlman and colleagues (2004) looked at the common variance accounted for by the factors, as well as the correlations between factors, and deduced that the ECERS-R measured one global

factor for classroom quality, citing as evidence for this claim the 71% of common variance accounted for by the first factor, and the correlations among factors which were > .50. However, the decision as to which factors to retain in the solution could have been driven by theoretical conceptualizations of classroom quality, as there was a theoretical alignment of the items onto their respective factors. Items that measured more of the process features of quality (e.g., such as interactions among teachers and students) all clearly reflected a distinct factor, while items measuring the presence of materials and activities in the classroom all distinctly reflected structural features of classroom quality. The factor analytic results were also limited by the selection criteria of the items that were utilized in the factor model, as researchers eliminated questions that lacked common variance (i.e., variance estimates of less than 0.10) or that contained "highly skewed" distributions. This selection criteria had the potential to impact the measurement properties of the instrument in several ways. First, the decision to eliminate items with little to no variance implied that, within the context of the larger measure, items which exhibited no variance were considered to provide no information about the underlying construct. However, it might have been the case that these items served as an important "floor" for the measure (i.e., were particularly easy for raters to endorse for each ECE center). These items might have acted as a risk indicator for classrooms with very poor levels of classroom quality. Second, ignoring the fact that the authors provided the reader with no objective criteria to understand how items were "highly skewed" to warrant exclusion from the analysis, items that were skewed might not have been problematic for measurement (Linacre, 1992; 1996). On average, we would expect that items that were more difficult for raters to endorse would demonstrate positive skewness, with only the programs with exceptional quality receiving higher

ratings. As such, the selection criteria might have actually decreased the range of classroom quality (i.e., the bandwidth) that the measure was capable of capturing.

Perlman and colleagues (2004) went further in their validity analyses, and explored another key issue that had been previously posited in the literature. In light of high inter-item correlations among items in the ECERS, Scarr, Eisenberg & Deater-Deckard (1994) had previously posited a shorter version of the measure would function as well as the full version of the measure. As such, Perlman and colleagues (2004) compared the alpha coefficients from: 12 randomly selected items from the measure, 10 items teachers chose as among the most difficult in the measure, and 24 items that teachers indicated were easy to administer. In all cases, the alphas of the scales were > .8, and correlations between these reduced versions of the measure and the total score were all > .8. Again, like Sakai et al. (2004), this study indicated that there might be some redundancies in the construct representativeness of the items contained in the measure; however, relying on alpha as a metric for the acceptability of shorter forms of the ECERS-R is misguided, and represents a critical deficit in the findings from research pertaining to the validity of the ECERS-R. The bandwidth fidelity issue in the design and modification of measures has been discussed by Singh (2004), whereby fidelity pertains to retaining items that are very similar to one another, and bandwidth pertains to selecting items that cover a wide range of the underlying continuum for a latent construct. Singh (2004) highlights how attempts to maximize alpha (i.e., the fidelity of the measure) always come at some expense to the bandwidth of a measure, as the range of the underlying continuum for the construct becomes restricted. Further, reliability coefficients like alpha provide no information about where, and what range, of the underlying latent continuum is captured by the items (i.e., the bandwidth). For example, a researcher could obtain a high reliability coefficient for a group of items that covered only a

relatively low level of classroom quality (i.e., low bandwidth). The ECERS-R purports to measure global classroom quality at multiple levels of the latent continuum, and in this hypothetical situation the high alpha would mask the deficiency in the measure. As such, it is clear that caution should be taken in interpreting the results of Perlman and colleagues (2004) with regard to strategies for abbreviating the ECERS-R. The similar alphas observed across each of their reduced versions of the measure might be indicative of construct saturation, but it is an inference that can only be made with respect to information about the bandwidth of the measure. To date, no validity investigations of the ECERS-R have explored issues concerning the bandwidth of the measure.

Cassidy, Hestenes, Hegde, Hestenes and Mims (2005a) also examined the dimensional structure of the ECERS-R using factor analytic techniques. Using a large sample (N = 1313 classrooms), the researchers were able to utilize cross-validation techniques to see whether the structure of the measure, as established with exploratory factor analysis, replicated across successive samples using confirmatory factor analysis. In doing so, the researchers demonstrated that a two-factor specification of the measure was adequate. Similar to previous validity studies, the two factors that emerged from this analysis pertained to language and interactions (i.e., Teaching and Interactions), and activities and materials available for children (i.e., Provisions for Learning). These dimensions appeared to represent distinct dimensions of quality, as the correlation between the two factors was of a moderate size (r =.46). Further, these results fit with theoretical conceptualizations of structural and process quality in two ways. First, the researchers suggested that the content of their items was theoretically aligned with distinct dimensions of structural and process quality. The researchers also showed that factor scores for each of these dimensions were moderately correlated with total scores for the ECERS-R, with correlations for

the two factors and the total score of the measure both > .70. Second, the results suggested that while the two factors might be related, they were not dependent on one another, meaning a classroom could be rich in resources, but poor in interactions, and theoretically low in quality, which provided weak support for the idea that the ECERS-R measures distinct aspects of quality. Further, the researchers were able to demonstrate that the factor scores for each dimension were more sensitive in distinguishing between the average score for the highest and lowest quantiles in the distribution of ECERS-R scores in the sample than compared to the total score for the measure. However, this study was flawed in a number of ways. Again, similar to prior validity analyses of the ECERS-R, the reliance on alpha as a useful metric for the quality of a measure is dubious at best (Sijtsma, 2009). Further, while it was useful to understand that the factor scores did not correlate to a high magnitude with the total score for the measure, that information provided little in the way of understanding whether there are classrooms that would have substantively different estimates for classroom quality on the dimensions when compared to their total score on the measure. This is an issue the field has yet to address, and it is critically important if both factors are to be considered as unique measures of classroom quality. Further, analyses showing that the factor scores were better able to distinguish between low and high levels of classroom quality have as an assumption that large differences in scores for classroom quality actually imply large differences in classroom quality. Again, the field has yet to provide any information about the range of the latent construct captured by the items from the measure. It could be the case that this assumption holds. However, it could also be the case that large differences in scores only distinguish between low-levels of classroom quality. Nevertheless, the factor solution proffered by Cassidy and colleagues (2005) does have strength, in that it replicates solutions found in prior research (Perlman et al., 2004; Sakai et al., 2004), and does so

using a robust sample of classrooms with cross-validation techniques. This has likely aided in this factor solution becoming the most widely utilized in the literature. As such, it deserves serious consideration as a unique specification of the measure, and should come with a full array of diagnostic psychometric information that can be used to argue for the validity of the use of these factor scores in both research and policy applications. However, to date, comprehensive psychometric information about this specification of the measure has been non-existent.

Cassidy, Hestenes, Hansen, Hedge, Shim and Hestenes (2005b) also described the dimensional structure of the ECERS-R using qualitative methodologies, and using a structural and process distinction. The researchers used a constant comparative analysis to identify whether each of the indicators for the rating scale for each of the 43 items measured structure or process quality. This study was notable in that it allowed for items to be associated with both structural and process dimensions (i.e., analogous to an item simultaneously loading onto two factors). The reason for this was because indicators that make up the rating scale for the ECERS-R could reflect mixtures of both process and structural quality. For example, the "Nature and Science" item contains indicators for low values of the rating scale that reflect the presence of activities and materials relating to nature and science in the classroom (i.e., structural quality), whereas high values of the rating scale contain indicators that correspond to how teachers interact with students with these activities and materials (i.e., process quality). Across the entire measure, the researchers found that 56% of items for the ECERS-R reflected structural quality, and 44% of items reflected process quality. This study is a unique departure from the previous approaches that have used factor analytic methods to describe the dimensionality of the measure given its use of qualitative techniques. Further, this specification of the measure does not force an item to reflect a specific dimension, and as such, might better reflect the true underlying structure of the

measure. To date, this specification of the measure has not been tested with quantitative methods. Other researchers have highlighted this mixing of both structural and process features within items as problematic for prior multidimensional specifications of the measure (Gordon, 2013; Lambert, Williams, Morrison, Samms-Vaughan, Mayfield & Thornberg, 2008), so there is reason to believe additional multidimensional specifications of the measure might be appropriate.

It is clear from the prior research that policymakers lack comprehensive diagnostic information about how the ECERS-R functions. This can be explained, in part, by an overreliance on using methods from classical test theory (CTT) like factor analytic techniques to explore the structural validity of the measure. The assumptions of these models can severely impact a user's understanding of the functioning of the measure. For example, the parameters from these models are completely sample dependent, whereby the observed score is item dependent, and the item statistics are sample dependent (Fan, 1998). As such, representative samples are important. Further, because the method seeks to obtain a minimum estimate of the number of factors that account for the inter-correlations between items, it favors retaining items that are similar to one another. This means, for measures like the ECERS-R, priority is given to retaining items that are nearly equivalent in terms of "endorsability." This leads to reductions in the range of the latent construct captured by the items (Singh, 2004), which is itself not an issue explored with these techniques. Consequently, there is ambiguity in how the population distribution of quality scores for the measure (or its sub-dimensions) overlaps with the range of classroom quality captured by the items. This can make it difficult to understand whether a statistically significant difference in the average ECERS-R scores between two groups actually aligns with substantive differences in quality. Missing data can also pose considerable challenges for these methods, as deleting items can alter the representations of the constructs, and as such,

alter the meaning of the relationships between items. Further, deleting persons alters the standardizing sample (Wright 1996, p.10). In addition, these methods assume that reliability is the same across different levels of the construct. However, this assumption could mask information related to the precision of the measure in different contexts, as it could be the case that the ECERS-R is most reliable only for a certain level of classroom quality. Interpretation of factor scores for classrooms is also not straight-forward, as these usually occur in reference to the sample mean. However, a score for classroom quality should encapsulate information about the kinds of items that define the current level of quality for a classroom, and the kinds of items that would need to be endorsed at higher levels to increase the quality of that classroom.

Modern psychometrics, such as item response theory, can be a remedy to the methodological shortcomings of CTT approaches. As such, scholars are increasingly calling for the widespread adoption of these methods across disciplines (Gordon, 2015a). IRT approaches can be contrasted with CTT in several ways. IRT methods comprise a family of models that consider a person's response to an item as a non-linear probabilistic function that is impacted by characteristics of the person and item. Because these models were first utilized to develop measures for mental ability, the nomenclature for person and item characteristic usually references the "ability" of a person and "difficulty" of an item. However, the basic idea easily generalizes. For example, in the case of the ECERS-R, "ability" could refer to the level of classroom quality, while "difficulty" could refer to the endorsability of the items. The logic of these models is quite straightforward: a classroom with a low-level of classroom quality is probably unlikely to be rated high for a difficult item. These models have several advantages over IRT methods. First, the raw score for a person is not assumed to be a linear measure, and is only linearized through a transformation of raw scores into logits or through proper fit of the

items to a 1-parameter Rasch model (Wright 1996, p.10). In addition, item parameters are independent of the sample used. For example, an item that is extremely difficult to endorse is likely to remain difficult in subsequent analyses. Further, items in the IRT approach are selected to cover a wide range of the dimension for underlying classroom quality. This can aid researchers in being explicit about the ordering of items along the underlying latent continuum for classroom quality. In addition, because the difficulty estimates for items are put on the same scale as the estimates for classroom quality, it is possible to make concrete comparisons about how the items overlap with the distribution of classroom quality scores. This can be useful in understanding on average how difficult or easy the pool of items was for the sample. It also can provide critical information about the construct saturation of items, which can aid developers in understanding which items could be removed from the measure. The same can be said for issues pertaining to construct inadequacy, as "holes" in the measure can easily be spotted along the continuum of classroom quality by regions not represented by items. Finally, because a reliability for each person is estimated, it is possible to examine where along the underlying latent continuum for classroom quality that the measure is most reliable. In sum, IRT approaches have the potential to provide a suite of diagnostic information to understand the functioning of the ECERS-R, which is necessary for making claims about the validity of the measure in practice (Kane, 2006).

Item Response Theory Analyses of the ECERS-R. Despite the use of the ECERS-R in research and practice over the last three decades, psychometric investigations of the measure using modern psychometrics are just now emerging. To date, only two studies have investigated the functioning of the ECERS-R using IRT methods and a sample of US preschools. Both of

these studies explicitly focused on understanding the functioning of the rating scale for the ECERS-R.

Gordon and colleagues (2013) conducted the first analysis of the ECERS-R using item-response theory. Using the Rasch Partial Credit model, the researchers investigated the response structure of the rating scale used for the ECERS-R using the ECLS-B dataset. The Partial Credit Modeling approach adopted by the researchers did not force ordering between the adjacent categories on the rating scale (Andrich, de Jong, & Sheridan, 1997), which allowed for explicit testing of the ordering between adjacent categories on the scale. The logic underlying their analysis was that, for any given item, a classroom with a higher-level of classroom quality should not be rated lower on that item than a classroom with lower quality. Instead, the researchers found that every item on the ECERS-R exhibited disordering of at least one category on the underlying rating scale. Given the stop-scoring routine recommended by developers of the ECERS-R, this analysis had important policy implications for how the measure is administered, as it implied that the stop-scoring routine likely led to inaccurate estimates for classroom quality. Further, while not explicitly mentioned by the researchers, disordered rating scales indicated that the total score for classroom quality were not valid interval measures of classroom quality, suitable for use in parametric statistics (Linacre, 2006).

Gordon and colleagues (2015b) followed up the 2013 investigation by examining the monotonic ordering of the indicators that define the categories for rating scale for every item on the ECERS-R. The rating scale for the ECERS-R asks observers to look for the presence of numerous indicators in the classroom. The presence of these indicators is attached to the rating scale, which asks observers to rate each item on a 7-point likert scale, whereby scores of 1 indicate inadequate quality and scores of 7 indicate excellent quality. However, as Gordon and

colleagues highlighted in their 2013 analysis, the stop-scoring routine can misrepresent estimates of classroom quality for each classroom when the rating scale is disordered. Similarly, the ordering of the indicators for the rating scale was also important to understand. As such, the researchers looked for monotonic functioning of the indicators that made up each category of the rating scale. Results for these analyses showed similar problems as their prior rating scale analysis, with indicators for categories of the rating scale disordered for most items on the ECERS-R. Further, the researchers were able to calibrate difficulty estimates for the indicators for each category of the rating scale, and showed that the ECERS-R was only capable of measuring low-levels of classroom quality.

Critical Gaps

Questions still remain about the internal structure and validity of the ECERS-R, and arguably the most useful IRT methods have yet to be utilized with this measure. For example, Gordon and colleagues (2013; 2015b) acknowledge that the Rasch models they employed were but one of many possible psychometric models that could be utilized for these investigations. For example, the Rasch methodology has been extended to accommodate multidimensional measures. The studies by Gordon and colleagues (2013; 2015b) assumed unidimensionality of the measure. In practice, there are many instances when the assumption of unidimensionality does not hold, and where the application of a multidimensional measurement model is both technically appropriate and substantively advantageous. Prior research has continually demonstrated a multidimensional structure for the ECERS-R. As such, there is reason to believe that the ECERS-R is not a global measure of classroom quality, but instead reflects unique domains of classroom quality, namely a Provisions for Learning and Teaching Interactions dimensions (Cassidy et al., 2005a; Perlman et al., 2004; Sakai et al., 2004) or Structural and

Process dimensions (Cassidy et al., 2005b). It is critical to understand the implications that these different dimensional specifications have for understanding classroom quality. In particular, information is needed to ascertain whether estimates for classroom quality for each of these dimensions could lead to substantively different conclusions about levels of quality for some classrooms when compared to their total score on the ECERS-R.

Further, the type of IRT model chosen to analyze the data should align with how the measure is used in practice. To date, most research and policy applications use total scores for either the full version of the ECERS-R or for the Provisions for Learning and Teaching and Interactions dimensions. These scores are only a valid measure of classroom quality if the data fits a particular kind of IRT model called the Andrich Rating Scale Model (1978), which is from the larger family of Rasch models. However, Gordon and colleagues (2013; 2015b) have shown, the rating scale for the measure is disordered. As such, total scores from either the full specification of the ECERS-R, or its multidimensional specifications, are not valid representations of underlying levels of classroom quality. However, researchers have posited ways to recover rating scales in instances where the categories are disordered, which is typically achieved through collapsing categories of the rating scale (Linacre, 2006). Given the widespread adoption of this measure in both policy and research contexts, it is critical that researchers look for ways to recover the rating scale, so that users can ensure the scores they are obtaining adequately represent true underlying levels of quality.

Another gap in IRT analyses involving the ECERS-R is that studies to date have provided no information about the fit of items to the underlying Rasch models. As such, all items were assumed to be valid measures of the classroom quality construct. However, the object of measurement is to discover the structure of quantity in the data. Within the context of Rasch

methodology this means identifying items from the measure that fit the underlying model, which might include trimming items from a measure that do not fit the underlying model. Further, this also means adequately defining the range of the underlying construct covered by the items. To this point, much work is still needed to investigate whether the ECERS-R contains redundant items, contains items that exhibit too much noise to contribute to productive measurement, or whether the range of the latent construct captured by the items is useful for meaningfully distinguishing different levels of quality between classrooms.

A more general critique of all prior validity studies is that none have embraced a comprehensive framework for organizing the psychometric properties of the ECERS-R. There are several assumptions underlying the use of observational measures like the ECERS-R in policy applications. Specifically, Charalambous et al. (2012, pg.3) have highlighted: (1) assumptions about the accuracy, reliability, and overall usefulness of the scoring mechanism; (2) assumptions about the generalizabilty of items in terms of their representation of the wider universe of possible items for classroom quality; (3) assumptions involving extrapolation, which typically focus around whether the assessment represents the constructs as intended, and the measure as a whole aligns with external domains of interest, and; (4) assumptions involving decisions, typically focused around whether consequences based on scores from the instrument are appropriate. To date, the evidence base for these assumptions is lacking (Mashburn, 2017; Pianta, 2012). Bringing new rigorous and comprehensive psychometric evidence to light to support these assumptions needs to be of central concern for any study investigating the psychometric properties of the ECERS-R. Attempts should be made to organize psychometric analyses of the ECERS-R within extant validity frameworks, as doing so will allow researchers and policymakers to understand the range of psychometric properties for the measure.

Messicks' Construct Validity Framework. Messick's unified concept of construct validity (1989; 1995) provides a means for categorizing the kinds of evidence needed to support arguments about the validity of measures. Messick (1989) describes validity as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (pg. 6). The framework moves away from the view of validity as being equal to some fixed value that exists for the researcher to uncover, and instead puts the emphasis on the actual and potential uses of scores derived from measures (Messick, 1995). In fact, the Standards for Educational and Psychological Testing, which were adopted by the American Education Research Association (AERA), American Psychological Association (APA) and the National Council on Measurement in Education (1999) adopted Messick's framework. The Standards state: "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests," and that "the process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations" (AERA, APA, & NCME, 1999. p.9). Messick documents six aspects of test validity that are necessary for understanding whether the interpretability of the scores derived from a measure are trustworthy and appropriate: (1) Content aspects of validity refers to the relevance and representativeness of the content of the items; (2) Substantive aspects of validity examine whether the response to the items are consistent with the theoretical rationales used to develop the content of items; (3) Structural aspects of validity assess the fidelity of the scoring structure to the structure of the construct domain; (4) Generalizability concepts of validity assess the degree to which the measures maintain their integrity across various contexts; (5) External concepts of validity examine the degree to which the measure correlates with similar

measures or diverges from dissimilar measures; (6) Consequential validity refers to the positive or negative social consequences of using a particular test in policy applications. The goal of psychometric analyses should be to provide empirical support for each of Messick's components of validity, which can then be used to create what Cronbach and Meehl (1955) describe as a nomological net of consistent, related empirical findings to support the intended use of the measure in varying contexts and for varying purposes.

Purpose of the Study

The goal of this dissertation is to provide a comprehensive account of the psychometric properties of the ECERS-R using Rasch modeling approaches. If data fits underlying Rasch models then it can be assumed estimates of classroom quality are linear, additive, interval-level, invariant, and hierarchical, and score from the measure can be used in parametric statistics. Further, several specifications of the measure have been posited in the literature, specifically a Provisions for Learning/Teaching and Interactions specification (Cassidy et al., 2005a; Perlman et al., 2004; Sakai et al., 2004), and a Structural/Process specification (Cassidy et al., 2005b). It is not clear which of these specifications provides the most theoretically-predicated, interpretable, reliable, and responsive estimates for classroom quality. The internal structure and measurement properties of an instrument can be directly impacted by the structural validity aspects of the measure. As such, there might be considerable tradeoffs in the measurement properties of the ECERS-R to consider if using multidimensional specifications of the measure. A primary aim of this dissertation is to examine the psychometric properties of each multidimensional specification of the measure that has been suggested in the literature. Results are organized using Wolfe & Smith's Rasch validity framework (2007), which posits specific Rasch tests that correspond to each component of Messick's validity framework. As such, results

from these analyses provide both researchers and policymakers with a range of psychometric information about the measure, which can be used as supporting information to assist or disprove the use of scores from the ECERS-R for its varied research and policy applications.

Psychometric Framework of This Study

This dissertation adopts a Rasch validity framework. Rasch models have increasingly been utilized in education research examining the functioning of measures for early childhood classroom quality (Colwell, Gordon, Fujimoto, Kaestner & Korenman, 2013; Gordon, 2013). The Rasch model is a probabilistic model that assumes the probability of a person's endorsing items is a function of their ability and the item difficulty. Rasch analyses of rating scale data can be used to construct interval-level data from polytomous items, provided the data fits the Rasch model. That is to say, the Rasch model is capable of linearizing a potentially underlying non-linear rating scale into interval-level data if the underlying items fit the Rasch model (Andrich, 1978). The family of Rasch models all meet the requirements for invariant measurement (Engelhard, 2013).

The goal of Rasch modeling is to derive a person's score on a measure that could be interpreted linearly on a continuum of the underlying construct. The measurement ideas underlying the Rasch model are fairly straight forward to follow. First, items aligning with a construct should be of the "same sort" (Rasch, 1960), which is to say items should reflect the same underlying construct. Second, the items should show variation along a continuum of difficulty from very easy to very hard (Rasch, 1960). Third, there should be no gaps along the continuum of item difficulty, as gaps in the continuum of item difficulty represent regions of the underlying construct that are not measured by the items (Rasch, 1960). Fourth, the underlying continuum of item difficulty should follow a logical and/or theoretically predicated progression

(Rasch, 1960). Fifth, items should discriminate equally between high and low performers on a measure in order for the measure to adequately represent classrooms with high and low levels of classroom quality (Rasch, 1960). Sixth, items should be independent. This assumption implies that responses to one question on a measure do not depend on the responses to any other question on the measure.

To estimate classroom quality scores and item difficulties, the Rasch model applies a natural log transformation to the matrix of responses to the items in the measure (Ludlow & Haley, 1995). The unit of measurement in a Rasch analysis is called the logit (i.e., log-odds unit). This unit of measurement typically ranges from -4 to +4 logits, though theoretically can expand to infinity. Rasch results expressed as logits have several desirable characteristics. First, the continuum of -4 to +4 provides a convenient metric for understanding how easy or hard items are to endorse on average, as items that have higher logit values are more difficult to endorse on average, whereas items that have negative logit values are easier to endorse on average. The same can be said for classrooms, as classrooms with higher logit scores have higher levels of quality than those with lower logit scores. Additionally, both the classroom quality scores and item difficulty calibrations from a Rasch analysis are on the same scale (i.e., the logit), which provides other desirable information about the measure and classrooms. For example, the utilization of Wright maps, which provide a graphical display of both the distribution of classroom quality scores and item difficulty calibrations, can provide valuable information about how the measure functions and how it might be improved. For example, these figures clearly show the item difficulty hierarchy, which can provide insights into the kinds of items that define low to high levels of classroom quality. In addition, these figures also demonstrate the granularity with which items measure the underlying latent continuum, which is dictated by the

closeness of items to one another along the latent continuum for classroom quality. Gaps in the measure, as defined by regions of the latent continuum that are not represented by items in the Wright maps, can provide insights into the kinds of items that might need to be developed to "fill-in" the measure. Additionally, regions of the underlying latent continuum where several items seem to measure the same location on the latent continuum might also indicate items that could be removed due to their redundancy with other items in the measure. Finally, the juxtaposition of the distribution(s) of the classroom quality scores against the items difficulty calibrations can also be informative, as it provides information about the targeting of the items to the classrooms in the sample. This can be instructive for understanding the potential use of a measure in policy efforts. For example, assume a policy intervention is interested in improving classroom quality from the fall to spring of a school year, and that the measure for classroom quality contains items representing the entire range of the underlying latent continuum for classroom quality. A group of classrooms with below average Rasch scores in the fall would be able to examine the items outside of the current range of their quality level to focus on the specific practices that would need to be implemented throughout the year to increase current levels of classroom quality.

Finally, as the field of psychometrics has progressed, multidimensional specifications of the Rasch model have been developed (Adams, Wu, Wilson, 2015; Briggs & Wilson, 2003). Because researchers have continually shown that the ECERS-R measures multiple dimensions of classroom quality (Cassidy et al., 2005a; Perlman et al., 2004; Sakai et al., 2004), there is reason to believe that multidimensional specifications of the Rasch model might be more suitable for analysis of the ECERS-R. The main distinction between unidimensional and multidimensional Rasch models is that several underlying latent traits are posited to influence the response

probability of persons (Reckase, 2009). The multidimensional specification of the Rasch model can be distinguished further between a within-item and between-item multidimensional specification. The within-item multidimensional specification allows for each item to simultaneously measure multiple dimensions, while the between-item multidimensional specification constrains the relationship between items and constructs so that each item only measures one specific dimension. As such, this flexibility in the multidimensional formulation of the Rasch model allows for examination of both multidimensional specifications of the ECERS-R that have been posited in the literature.

Research Questions

Content Validity. Do the items from the ECERS-R exhibit a similar technical quality across different specifications of the measure (i.e., do the items fit an underlying Rasch model that would support the use of the total scores in practice)? Given that the ECERS-R was developed from best practices in early childhood research, it was hypothesized that items from the ECERS-R would exhibit few issues with fit to the underlying Rasch model. Further, prior research has suggested that a version of the ECERS-R containing a smaller number of items would be sufficient for measurement (Scarr et al., 1994; Perlman et al., 2004). As such, it was hypothesized that specifications of the ECERS-R which used a subset of items from the entire measure would not exhibit worse fit than the full 37-item specification of the measure.

What adjustments are necessary to the rating scale in order to allow for total scores to accurately represent underlying levels of classroom quality? Issues regarding the monotonic functioning of the rating scale are known (Gordon et al., 2013; 2015b). As such, it was hypothesized that these issues would be replicated using additional models from the larger family of Rasch models. Because the rating scale for the ECERS-R only defines the odd portions

of the rating scale with indicators (i.e., 1, 3, 5, and 7; Harms et al., 1998), it was hypothesized that a 4-point rating scale, which collapses the even categories of the rating scale into the odd categories, would function appropriately (i.e., monotonically) for measurement.

Substantive Validity. Is the item difficulty hierarchy for each specification of the measure supported by the developmental theory? Prior research has indicated that items from the measure which focus on interactions between children and teachers in the classroom are likely to be associated with higher levels of quality (Perlman et al., 2004; Pianta et al., 2005). As a result, it was hypothesized that the interactional items from the measure would be the most difficult items across all specifications of the measure.

Generalizability Validity. What are the reliability estimates from the various dimensional specifications of the measure? In line with prior studies that have reported reliability estimates for the various factors in the measure (Cassidy, et al., 2005a; Clifford et al., 2005b; Gordon et al., 2013; Perlman et al., 2004; Sakai et al., 2003), it was expected that the Rasch Person Reliabilities would all be greater than .70.

What evidence is there for item bias in the measure? Prior research has shown that teachers who hold a graduate degree are rated higher on measures of classroom quality (Pianta et al., 2005). Further, there is emerging evidence that Head Start centers score higher on the ECERS-R than other modes of caregiving on measures of classroom quality (Coley et al., 2014). As a result, it was hypothesized that, if item bias is present in the measure, it would privilege (i.e., items would be easier for) teachers with a graduate degree and Head Start centers.

Structural Validity. Which dimensional specification of the ECERS-R functions best to measure classroom quality? Further, do the different multidimensional specifications of the measure lead to substantive differences in estimates of classroom quality? In line with (Cassidy,

et al., 2005a; Clifford et al., 2005; Gordon et al., 2013; Perlman et al., 2004; Sakai et al., 2003) it was expected that the Cassidy et al. (2005a) Provision for Learning/Teaching and Interactions specification of the measure would best fit the data. In addition, several researchers have indicated that items from the ECERS-R contain mixtures of interactional and structural components of quality (Cassidy et al. 2005b; Gordon et al., 2015b; Perlman et al., 2004). As a result, it was hypothesized that the Cassidy et al. (2005b) Structural/Process within-item dimensional specification of the measure would also adequately fit the data, and meaningfully measure multiple dimensions of classroom quality.

External Validity. How many statistically distinct levels of classroom quality can the ECERS-R measure? The developers of the ECERS-R imply that the measure is capable of measuring four distinct levels of classroom quality (Harms et al., 1998). However, other researchers who have investigated thresholds for quality with the ECERS-R have found two distinct levels of quality (Abner et al., 2013; Burchinal et al., 2011; Le at al., 2015). As a result, it was hypothesized that the different specifications of the measure would be able to capture between two to four statistically distinct levels of classroom quality.

How do the different dimensional specifications of the measure relate to the Arnett Caregiver Interaction Scale (ACIS), a similar measure of quality? Given the alignment between content of the ACIS and the interactional items on the ECERS-R, it was hypothesized that the Process and Teaching and Interactions dimensions would show stronger associations with the ACIS than the other dimensional specifications of the measure.

Predictive Validity. What are the associations between the different dimensional specifications of the measure and preschool children's reading and math outcomes? Numerous prior studies have shown either no associations between the measure and child outcomes or small

associations with very modest effect sizes. As a result, it was hypothesized that no dimensional specification of the measure would relate to child outcomes.

Method

Data

Data was drawn from the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) (Flanagan & West, 2005). The ECLS-B is a multi-source, multi-method study and nationally representative sample that focuses on the early home and educational experiences of children. It was designed to collect data about the care and education, health, and development of children from birth through kindergarten entry. The study's primary sponsor is the National Center for Education Statistics (NCES) within the Institute of Education Sciences (IES) of the U.S. Department of Education. The sampling frame for the study was selected from the 2001 birth records of children in 46 states. The survey had a 74% response rate, and resulted in a sample of about 14,000 children at baseline. Fifty-one percent of these study children were boys; 54% were non-Hispanic White, 26% were Hispanic, 14% were non-Hispanic African American, 3% were Asian/Pacific-Islander, and 4% were of other race-ethnicities. At the 4-year-old follow-up interview, the sample size was 8,950, reflecting the exclusion of children who had died or had moved permanently out of the country, as well as those who could not be located, refused to participate, or lived more than 150 miles from the nearest interviewer (U.S. Department of Education, National Center for Education Statistics, 2001).

Sample Characteristics and Sample Size. The total available sample size of preschool centers that had item level information for the ECERS-R available was N = 1400 classrooms[1]. This was the sample size for both the psychometric and student outcome analyses that follow[2].

[1] The sample size has been rounded to the nearest 50 per the ECLS-B data reporting requirements.
[2] It might have been expected that the sample size for the child outcome analyses would be larger; however, the sample size was identical for both sets of analyses because individual children were followed throughout the data collection efforts for the ECLS-B, and as a result multiple children were not nested inside a given classroom.

The average age of children in the sample was 52.81 months (SD = 3.90) [3]. The majority of the children in the sample were male (53.69%). In addition, the majority of the children were also White (50.30%), followed by Black (20.27%), Hispanic (15%), and other reported race (6.42%). The average age of mothers in the sample was 32.03 years (SD = .86). The occupational status of mothers in the sample showed that 46.51% worked 35 or more hours a week, 19.30% worked less than 35 hours a week, 7.11% were looking for work, 25.26% were not in the workforce, and 1.62% provided no information about their work status. The average score on the composite measure for socioeconomic status (SES)[4] was .12 (SD = .86), which indicated that the sample was slightly above average in SES. Finally, the average number of children in each household that were under the age of 18 was 2.45 (SD = 1.16).

The majority of caregivers in the sample were female (98.20%), with male (1.79%), and not reporting (.01%). Most of the caregivers were White (65.75), with Black (18.92), Latino (4.76), and other reported race (10.57%). Teachers' education statuses were: less than high school (2.1%), high school (7.65%), vocational school (1.94%), some college (16.05%), Associate's degree (18.52%), Bachelor's degree (34.94%), and Graduate degree (18.81). Most center locations were located in Urban areas (70.77%), followed by suburban locations (12.78%), and rural locations (16.46%). Less than half of the sample contained teachers who were teaching under 30 hours a week (46.71%). Finally, the center types in the sample consisted of Head Start (9.17%), private (2.2%), State/Local (27.53%), and other center types (61.11%).

[3] All descriptive information was calculated using the W31CO weight.
[4] Each ECLS data file includes a composite measure of SES reflecting the SES of a child's household at the time of data collection. The components used to create the SES variable are father/male guardian's education, mother/female guardian's education, father/male guardian's occupational prestige, mother/female guardian's occupational prestige, and household income. In households with two mothers or two fathers, education and occupational prestige for both mothers/fathers were used (Flanagan & West, 2005).

Measures

Child Assessment Data. Two composite scores for children's reading and math skills were used as measures of child outcomes (i.e., X3RTHR2 and X3MTHR2, respectively). These scores were created using item response theory by creators of the ECLS-B, and were based on items from several extant measures for children's cognitive skills. The IRT-developed reading composite contained items derived from the following measures: the three subtests of the Preschool Language Assessment Scales (Simon Says, Art Show and Let's Tell Stories; Duncan & De Avila, 1998), the Peabody Picture Vocabulary Test–Third Edition (Dunn & Dunn, 1997), as well as other items which were designed to measure letter sounds, early reading, phonological awareness, knowledge of print conventions, and matching words. The overall reliability of the IRT ability estimate for the reading composite was .84 (a comprehensive account of the procedures used to calculate IRT reading ability estimates and reliability statistics can be found in Najarian, Snow, Lennon, Kinsey & Mulligan, 2007, pp. 77-84). Items that constituted the math composite were derived from the following measures: Test of Early Mathematics Ability (Ginsburg & Baroody, 1983), as well as other sources pertaining to number sense, geometry, counting, operations, and patterns (Najarian, et al., 2007). The overall reliability of the IRT ability estimate for the math composite was .89 (information about the procedures used to calculate IRT math ability estimates and reliability statistics can be found in Najarian, et al., 2007, pp. 85-95). The composites created from the ECLS-B assessment data are reliable indicators of children's reading and math skills that have been used in countless studies to date (see Najarian, et al., 2007).

Measure of Classroom Quality. The 37 items for the ECERS-R formed the basis for the Rasch analyses of the measure. The ECERS-R contains items that are conceptually grouped

under seven subscales (i.e., Space and Furnishings, Personal Care Routines, Language-Reasoning, Activities, Interaction, Program Structure, and Parents and Staff). Developers of the ECLS-B chose not to administer all 43 items of the ECERS-R because of overlap between some items from the ECERS-R and other sources of data collected in the study. In particular, developers of the ECLS-B dataset chose to omit the items from the Parents and Staff sub-scale. Items from the ECERS-R are scored on a 7-point rating scale, with categories from the rating scale described in the following way: 1 (inadequate quality), 3 (minimal quality), 5 (good quality), and 7 (excellent quality). For most items in the ECERS-R trained observers look for the presence of specific indicators in the classroom, and these indicators corresponded to the categories of the rating scale described above. However, some indicators for items require the staff to answer response prompts from the observer. Finally, developers of ECER-R recommend a single three-hour observation period to complete the measure. The percentage of responses for each category of the rating scale for classrooms in the present sample can be found in Table 2 below[5], while the weighted means and standard deviations for the items are found in Appendix A.

Table 2. *Percentage of Quality Response for Each Category of Rating the Scale*

Item Name and Number	Subscale	%Cat 1	%Cat 2	%Cat 3	%Cat 4	%Cat 5	%Cat 6	%Cat 7
1 Indoor Space	Space and Furnishings	4	5	6	28	4	13	39
2 Furnishings for routine care	Space and Furnishings	2	1	0	7	3	23	64
3 Furnishings for relaxation	Space and Furnishings	7	5	19	34	7	10	18
4 Room Arrangement	Space and Furnishings	4	7	9	16	6	13	46
5 Space for Privacy	Space and Furnishings	7	6	24	29	6	10	19
6 Display for Children	Space and Furnishings	2	9	24	29	10	16	11
7 Gross Motor Space	Space and Furnishings	8	26	7	20	11	15	12
8 Gross Motor Equipment	Space and Furnishings	9	20	5	17	5	17	26
9 Greeting/Departing	Personal Care Routines	3	5	4	12	3	9	64

[5] The Parents and Staff items from the ECERS-R are not included here because there were not collected as part of the ECLS-B study.

Table 2. *Percentage of Quality Response for Each Category of Rating the Scale*

Item Name and Number	Subscale	%Cat 1	%Cat 2	%Cat 3	%Cat 4	%Cat 5	%Cat 6	%Cat 7
10 Meals/Snacks	Personal Care Routines	40	18	1	9	6	10	16
11 Nap	Personal Care Routines	12	20	3	22	2	3	9
12 Diapering/Toileting	Personal Care Routines	30	24	1	11	1	9	24
13 Health Practice	Personal Care Routines	6	51	1	7	3	11	21
14 Safety Practice	Personal Care Routines	26	22	1	10	2	7	32
15 Books and Pictures	Language and Reasoning	3	6	7	59	2	3	20
16 Encouraging to Communicate	Language and Reasoning	2	3	5	20	6	23	42
17 Language to develop reasoning	Language and Reasoning	6	6	20	29	6	9	26
18 Informal use of language	Language and Reasoning	2	2	8	28	3	13	43
19 Fine motor activities	Activities	2	6	10	42	4	12	26
20 Art	Activities	5	9	21	38	4	10	13
21 Music and Movement	Activities	3	28	14	32	8	7	7
22 Blocks	Activities	8	8	5	48	8	19	4
23 Sand and water play	Activities	16	3	17	29	7	16	12
24 Dramatic Play	Activities	6	10	12	49	11	11	3
25 Nature and Science	Activities	17	22	8	39	1	3	9
26 Math	Activities	8	3	13	53	4	6	12
27 Use of TV, video or computer	Activities	12	20	2	18	5	10	9
28 Promoting acceptance of diversity	Activities	5	7	23	33	11	8	13
29 Gross motor supervision	Interaction	8	5	5	20	19	19	24
30 General supervision	Interaction	7	6	2	13	11	16	45
31 Discipline	Interaction	4	6	4	13	15	24	34
32 Staff-child interactions	Interaction	4	5	2	10	1	7	70
33 Interactions among children	Interaction	3	5	3	14	3	27	47
34 Schedule of daily play	Program Structure	4	27	2	28	2	8	29
35 Free play	Program Structure	5	7	6	27	5	14	36
36 Group time	Program Structure	7	2	6	14	5	17	50
37 Provisions for exceptional children	Program Structure	1	4	1	3	1	6	21

Analytic Methods

Analytic Summary of Rasch Methodology. In order to answer each research question, Wolfe and Smith's (2007) Rasch validity framework was utilized. The framework prescribes specific Rasch tests that align with each component of Messick's (1989) validity framework. These tests were carried out for four different specifications of the ECERS-R. First, a

unidimensional specification of the ECERS-R was examined, whereby all 37 items for the ECERS-R that were available in the ECLS-B dataset, were specified to measure a single unidimensional (i.e., "global") construct. Next, the Cassidy and colleagues (2005a) Provisions for Learning/Teaching and Interactions multidimensional specification of the measure was examined. A complete list of items that were used in this specification of the measure is found in Table 3. However, this specification of the measure only utilized 16 items from the measure. As such, it was not directly comparable to the full 37-item unidimensional specification. Consequently, in order to aid in comparing "apples to apples", the 16 items used in the Provisions for Learning/Teaching and Interactions specification of the measure were also examined as a separate unidimensional measure for quality.

Table 3. *Items from the ECERS-R Corresponding to the Provisions for Learning/Teaching and Interactions Specification of the ECERS-R*

Item	Dimension
(3) Furnishing for Relaxation	Materials/Activities (i.e. Provisions for Learning)
(5) Space for Privacy	Materials/Activities (i.e. Provisions for Learning)
(15) Books and Pictures	Materials/Activities (i.e. Provisions for Learning)
(1) Fine Motor	Materials/Activities (i.e. Provisions for Learning)
(20) Art	Materials/Activities (i.e. Provisions for Learning)
(22) Blocks	Materials/Activities (i.e. Provisions for Learning)
(24) Dramatic Play	Materials/Activities (i.e. Provisions for Learning)
(25) Nature/Science	Materials/Activities (i.e. Provisions for Learning)
(26) Math/Number	Materials/Activities (i.e. Provisions for Learning)
(17) Using Language to Develop Reasoning Skills	Teaching and Interactions
(18) Informal Use of Language	Teaching and Interactions
(30) General Supervision of Children	Teaching and Interactions
(31) Discipline	Teaching and Interactions
(32) Staff-child Interactions	Teaching and Interactions
(33) Interactions among Children	Teaching and Interactions
(36) Group Time	Teaching and Interactions

Note. Adapted from Cassidy et al (2005[a]). The items used in this analysis were also used in the 16-item unidimensional specification.

Finally, the Structural/Process specification of the measure proffered by Cassidy and colleagues (2005b) was examined. This specification of the measure allowed for some items to simultaneously measure both Structural and Process dimensions of the measure. A complete list of the items and their corresponding dimension(s) is found in Table 4. Again, because the goal

was to directly compare these different specifications of the measure, only the 16 items that were used in the Cassidy and colleagues (2005a) Provisions for Learning/Teaching and Interactions specification were used for these analyses. Finally, a graphical display of the differences in how the dimensional specifications of the measure differed can be found in Figure 2.

Table 4. *Items From the ECERS-R Corresponding to the Structural and Process Dimensions*

Item	Dimension(s)
(3) Furnishing for Relaxation	Structural
(5) Space for Privacy	Structural/Process
(15) Books and Pictures	Structural/Process
(1) Fine Motor	Structural
(20) Art	Structural/Process
(22) Blocks	Structural
(24) Dramatic Play	Structural
(25) Nature/Science	Structural/Process
(26) Math/Number	Structural/Process
(17) Using Language to Develop Reasoning Skills	Process
(18) Informal Use of Language	Process
(30) General Supervision of Children	Process
(31) Discipline	Structural/Process
(32) Staff-child Interactions	Process
(33) Interactions among Children	Process
(36) Group Time	Structural/Process

Note. Items denoted as "Structural/Process" were items that were specified to provide information to both the Structural and Process dimensions using the multidimensional Rasch analytic procedures.

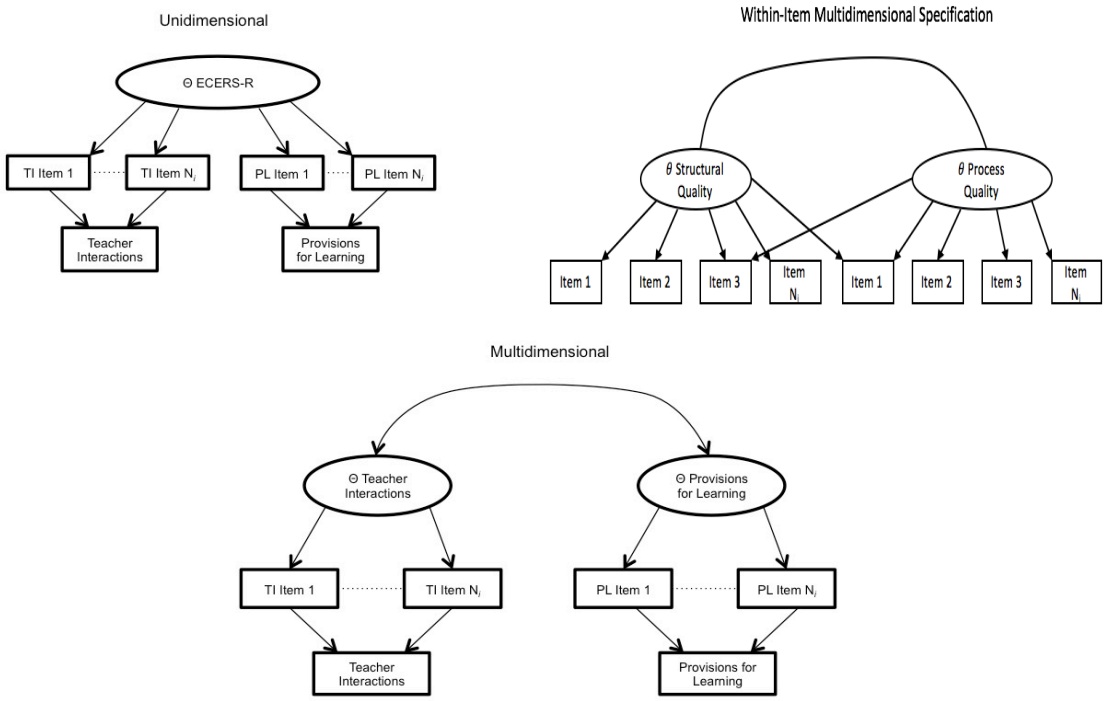

Figure 2. Graphical depictions of the different ways to model multidimensionality using the proposed Rasch analytic procedures.

Psychometric modeling was conducted using the TAM package (Kiefer, Robitzsch & Wu, 2015) in R version 3.3.2 Sincere Pumpkin Patch (R Core Team, 2016). The Andrich (1978) Rating Scale model was used to examine the psychometric properties of the two unidimensional specifications of the ECERS-R (i.e., 37-item and 16-item). The model assumed that the response categories had the same meaning across all items. The equation for this model was as follows:

$$\pi_{nix} = \frac{e^{\sum_{j=0}^{x_{ni}}\left[\beta_n - \left(\delta_i + \tau_j\right)\right]}}{\sum_{k=0}^{m} e^{\sum_{j=0}^{x_{ni}}\left[\beta_n - \left(\delta_i + \tau_j\right)\right]}} \quad (1)$$

π_{nix} was the probability of classroom *n* being rated in category *x* for item *I;* δ_i was the location of the item difficulty calibration for item *i* on the underlying latent continua of classroom quality; τ_j was the location of the k^{th} transition from on response category to the next for the *m*+1 rating

categories; while β_n was the parameter for a classroom's level of quality (i.e., sometimes colloquially referred to as a "Rasch score"). Both multidimensional specifications of the measure (i.e., Provisions for Learning/Teaching and Interactions and Structural/Process specifications) were fit using a multidimensional random coefficients multinomial logit model (MRCMLM; Adams, et al., 2015; Briggs & Wilson, 2003). The equation for the MRCMLM was as follows:

$$\mathrm{P}(\mathrm{X}_{ik} = 1; \mathbf{A}, \mathbf{B}, \xi|\theta) = \frac{\exp(\mathbf{b}_{ik}\theta + \mathbf{a}'_{ik}\xi)}{\sum_{k=1}^{K_i} \exp(\mathbf{b}_{ik}\theta + \mathbf{a}'_{ik}\xi)} \quad (2)$$

Matrices **A** and **B** were the scoring and design matrices, and were used to specify the functional form of the model relative to the hypothesized mapping of items to dimension (**A**) and the item difficulty calibrations (**B**), and allowed for specifying either the within- or between-item dimensional specifications of the ECERS-R; the responses to the measure were modeled as a linear function of the underlying level of the latent trait θ on the dimensions and relative to the difficulty of category k($\mathbf{b}_{ij}$) for an item; K_i was the number of categories for item *i* (for the ECERS-R K_i = 7 for every item); the vector of location parameters for the items was represented by ξ; $\mathbf{b}_{ik}$ was the scoring vector for category *k* of item *i* across the latent dimensions; and $\mathbf{a}_{ik}$ was the design vector given to category *k* of item *i* that described the linear relationship among the elements of the vector for the location parameters.

Wolfe & Smith's (2007) Validity Tests. Wolfe and Smith (2007) provide a comprehensive account of a series of instrument development activities, using Rasch methods, which are necessary to provide validity evidence for each component of Messick's (1989) construct validity framework. These activities were undertaken for each of the four specifications of the measure that were examined. Each of these actives are described in detail below, and a general overview of all the activities can be found in Table 5.

Table 5. *Adaptation of Wolfe & Smith's (2007) Conceptualization of Rasch Validity Evidence for Messick's Validity Framework*

Content	Substantive	Generalizability
Item Technical Quality	Rating Scale Functioning	Item Difficulty Invariance
	Item Difficulty Hierarchy	Differential Item Functioning
		Reliability of the Estimates for Classroom Quality (i.e., "Rasch Scores)
		Precision of Estimates for Classroom Quality
Structural	**External**	**Predictive**
Goodness of Model Fit	Rasch Person Strata Indices	Regressions with Child Outcomes
Comparison of Rasch Model Subscale Correlations	Coverage of Item Difficulty Calibrations	
Discrepant Case Analyses	Wright Maps	
	Correlations with Arnett Caregiver Interaction Scale	

Content Validity. The Rasch Infit and Outfit statistics were used to examine the fit of each item to the their respective Rasch models. Both the Infit and Outift mean squared error statistics described the differences between the observed responses and expected responses (i.e., the person-by-item residuals; Wright & Stone, 1979). The Infit statistic, which was calculated using a weighted average of the squared residuals, was useful for examining the consistency in responses across all items. The Outfit statistic, which was calculated through taking the average of the squared residuals, and was useful for examining whether classrooms who scored unusually low or high provided unexpected responses to items. Both the Infit and Outfit statistics have an expectation of one. Values less than one indicated overfit to the model, whereas values greater than one indicated underfit to the model. Underfit of the items to the model meant that items were less predictable than what would have been expected, and indicated a deficiency in the

model predicted randomness[6]. Data which overfit the model indicated that items were more predictable than what would have been expected by the model. Items with both Infit and Outfit statistics that fell between .7 and 1.3 were considered productive for measurement (Linacre, 2002)[7].

Substantive Validity. The rating scale provides substantive validity evidence when responses on the rating scale are consistent with the intentions of the instrument developers. The primary assumption of the measure is that the average classroom quality estimates increase with values of the 7-point rating scale, which is described as monotonic functioning. This is critical to investigate in order to ensure that total scores from the measure are valid measures of underlying levels of classroom quality. In line with Linacre (2010), the rating scale of the measure was investigated for monotonic functioning in the average classroom quality estimates for each category of the rating scale. Categories that failed to function monotonically were collapsed into adjacent categories until monotonic functioning of the rating scale was achieved. Further, the item difficulty calibrations were examined to assess whether the hierarchy of item difficulty conformed to theoretical conceptualizations for the ordering of item difficulty in relation to the content of the items. As such, it was expected that items pertaining to interactions between children and teachers would be among the most difficult in the measure. In addition, the average classroom quality estimates were also compared across all specifications of the measure to

[6] When data underfits the Rasch model it indicates a deficiency in the model predicted randomness. The deficiency can be calculated with $1 - Infit$ or $1 - Outfit$. For example, Infit = 1.65 is a 65% deficiency in model predicted randomness. This indicates there is 65% more noise in the data than was modeled.

[7] Wright, B.D. & Linacre, J.M (1994) offer several plausible ranges of values that are suitable for productive measurement, and propose that the appropriateness of any given range of values depends on the intended use of a test/measure in practice (e.g., high-stakes, run of the mill, survey, and clinical observation). These different ranges are found in the following publication: Reasonable mean-square fit values. *Rasch Measurement Transactions*, *(8)*3 p.370.

explore whether the scores for classroom quality were in line with theoretical conceptualizations for item difficulty.

Generalizability Validity. Several tests were carried out to examine the degree to which the measure maintained its integrity across different measurement contexts. First, correlations between the item difficulty calibrations for pairs of items were examined across each specification of the measure. Invariance of the item difficulty calibrations was established if the calibrations were strongly correlated across the different specifications of the measure, which would imply that the difficulty calibrations of the items were not sensitive to the different specifications of the measure. Next, each item was examined for item bias through generating estimates for differential item functioning (DIF). Item bias could have occurred if classrooms with identical levels of quality had different probabilities of endorsing an item based on group membership. Item bias was investigated for the following variables: center type, teacher education and half time program status. Estimates for significant DIF has been an ongoing topic in the literature (Hambleton, 2006; Zwick, Thayer & Lewis, 1999), but this study settled on the conventional standard of DIF $\leq$ -.5 or $\geq$ .5 as indicative of items that exhibited item bias (Wright, & Panchapakesan, 1969). In addition, the reliability of the estimates for classroom quality were also compared across each specification of the measure. As such, the estimates for the Expected *a posteriori* (EAP) reliability coefficients were examined across all specifications of the measure. Further, because reliability could change as a function of the underlying level of classroom quality, precision plots were also examined. Precision plots provided a graphical depiction of the relationship between standard errors and estimates of classroom quality. In particular, because test information is the inverse of the variance associated with each estimate of classroom quality, examining the range of estimated classroom quality with the smallest standard

errors was useful for understanding where the along the latent continuum for classroom quality that information from the measure was maximized. In addition, comparing precision plots across the different measure specifications was useful for understanding which specifications of the measure were associated with the most test information.

Structural Validity. Several analyses were carried out to explore the dimensional structure of the measure. First, the goodness of fit statistics were examined for each specification of the measure. In particular, the Akaike information criterion (AIC), Bayesian information criteria (BIC), sample adjusted Bayesian information criteria (aBIC), bias-corrected AIC (AICc), and the Bozdogan's Consistent Akaike information criterion (CAIC) were all utilized. The goal with this analysis was to establish which model exhibited the smallest model fit statistics. Second, the correlations between the estimates for classroom quality were examined in order to establish whether they were positively associated. Of particular interest were the correlations between pairs of dimensions for the Provisions for Learning/Teaching Interactions specification, and the Structural/Process specification of the measure. In order for multidimensional specifications of the measure to be warranted, correlations between dimensions needed to be positive, but not to a degree that would warrant a simple unidimensional specification of the measure. Finally, discrepant case analyses were undertaken to investigate whether multidimensional specifications of the ECERS-R led to substantive differences in estimates for quality (Briggs & Wilson, 2003). In order to establish discrepant cases in the sample, two sets of discrepant case analyses were conducted. In the first set of discrepant case analyses, each classroom quality estimate was standardized, and differences between dimensional estimates and the scaled estimates for both unidimensional (i.e., 37-item and 16-item) classroom quality estimates were examined. The goal was to establish the cases which differed by more than one

standard deviation on any dimensional estimate for quality when compared to their unidimensional quality estimates (Allen & Wilson, 2006). The second set of analyses calculated the sums of squares discrepancy indicator, DI_p^2 (Allen & Wilson, 2006). This was a metric helped to establish the cases that differed notably for the combined dimensional estimates for each multidimensional specification of the measure.

External Validity. Evidence for the external validity of the measure entailed documenting the responsiveness of the measure. To accomplish this the Rasch Person Separation (**G**) and Person Strata (**H**) indices were examined. Both indices provided information about the number of statistically different classroom quality performance strata that the measure could identify in the sample (Andrich, 1982; Wright & Masters, 1982)[8]. The Wright maps were then examined to inspect the relationship between the distributions of the item calibrations in relation to the classroom quality estimates. These maps were useful for understanding a number of different aspects which could influence the responsiveness of the measure, primarily because these maps display both the item calibrations and estimated classroom quality levels on the same scale. As a result, viewing these maps made it possible to establish whether the classroom quality estimates exceeded the average item difficulty calibrations, which would indicate that the measure was too easy for the classrooms in the sample. To aid in the interpretation, coverage statistics were calculated, which detail the percent of classrooms with estimates for quality within the range of the item difficulty calibrations. In addition, the spread of the item calibrations was also examined across the entire range of the latent continua, which made it possible to spot holes in the measure (i.e., regions of the latent continua not covered by item calibrations), and to spot regions of the

[8] The **G** index is a conservative statistic to determine the number or error strata (i.e., performance levels) in the sample that is appropriate when outliers in a sample are normally distributed, while the **H** index is sensitive to outliers in the sample that occurred as a result of extreme performance levels.

latent continua that might contain redundant items (i.e., item calibrations that overlap). Finally, it was also important to examine whether the Rasch generated estimates for classroom quality were concurrent with similar measures. As such, associations between the Rasch-generated classroom quality estimates and the total score from the Arnett Caregiver Interactions Scale were examined using Pearson correlation coefficients. It was expected that the correlations between all of the classroom quality estimates and the Arnett Caregiver Interactions Scale would be positive, with the strongest associations between the Teaching and Interactions and Process dimensions followed by the Provisions for Learning and Structural dimensions.

Predictive Validity. In the final step of the validity analyses, the associations between the Rasch-generated estimates of classroom quality and child outcomes were explored through a series of weighted multiple regressions using the *svy* package in Stata version 14 (StataCorp, 2015). The W31C0 survey weight was applied to the estimation of parameters in order to account for the complex survey design of the ECLS-B, and to ensure standard errors for the parameter estimates were unbiased. Regressions were carried out of each specifications of the measure. The independent variables used in the models can be found in Table 6 below. All continuous explanatory variables were grand-mean centered, so that the zero values for the intercepts and the continuous independent variables could be interpreted meaningfully.

Table 6. *Variables Used in the Multiple Regression for All Models*

Child/Family Covariates
Child Gender
Child Race
Child Age
Child Hispanic Identification
Mother's Highest Level of Education
Mother's Employment Status
Mother's Age
Received WIC Within the Last 12 Months

Table 6. *Variables Used in the Multiple Regression for All Models*

Number of children living in the House Under 18 Years of Age
ECLS-B Composite for Family SES Status
Provider Covariates
Provider Education Level
Provider Race
Provider Gender
Center Type
Center Location (i.e., Urbanicity)
ECERS-R Rasch Scores
Outcomes
ECLS-B Reading Composite
ECLS-B Math Composite

Note. N = 950 for cases with complete data on all of the covariates. The sample size has been rounded to 50 in line with the ECLS-B user agreement.

The primary purposes of these analyses were to establish whether the different Rasch generated estimates for classroom quality were significantly associated with child outcomes, and to establish whether the estimates for classroom quality accounted for the variance in student outcomes over and above the other variables in the model. In order to isolate the variance in student outcomes accounted for by the classroom quality estimates, the covariates were staged into the model in two blocks. In the first block, all the variables from Table 6, except for the classroom quality estimates, were staged into the model to establish a baseline R^2 values. Next, the classroom quality estimates were staged into the model. In the case of the two multidimensional specifications of the measure, scores from the two dimensions were entered into the model simultaneously. In the event that relationships between classroom quality estimates and student outcomes were statistically different than 0, the change in R^2 was examined to establish whether the ECERS-R scores meaningfully contributed to the overall fit of the models. Effect sizes for these models were also calculated using Cohen's f^2.

In order to examine whether missing data would pose a problem for these analyses, *a priori* power analyses were conducted using GPower version 3.1 (2009). In the case of the two

unidimensional specifications of the measures, these analyses assumed f^2 =.10, α = .05, with 35 predictors in the model, and a desired power of .80. Results showed that a samples size of N = 82 would be sufficient to detect an effect size as small as .10. Both of the multidimensional specifications of the measure assumed f^2 =.10, α = .05, with 36 predictors in the model, and a desired power of .80. Results showed that a samples size of N = 83 would be sufficient. In both cases results showed that the forthcoming analyses were sufficiently powered to detect minute associations between classroom quality Rasch scores and child outcomes.

Results

Content and Substantive Validity Evidence

In order to provide both content and substantive validity evidence for each specification of the ECERS-R, a set of analyses was carried out to: assess the rating scale functioning, examine the technical quality of the items, and to examine the item difficulty hierarchy across model specifications. The goal of these analyses was to answer the following questions: (1) Are technical quality of items similar (i.e., invariant) across different specifications of the measure?; (2) What revisions to the rating scale are necessary to support the use of total scores in both research and policy applications?; (3) Are the item difficulty calibrations invariant across different specifications of the measure?; and (4) is the item difficulty hierarchy in line with what has been proposed in the literature?

Item Technical Quality. In order investigate whether the technical quality of the items was invariant across different specifications of the measure, item misfit was examined using both the Rasch Outfit and Infit statistics. The Outfit and Infit statistics are reported for each specification of the measure in Tables 7 and 8, respectively.

Table 7. *Rasch Outfit Statistics for Each Specification of the Measure*

Items	Subscale	$Outfit^{37}$	$Outfit^{16}$	$Outfit^{PT}$	$Outfit^{SP}$
1 Indoor Space	Space and Furnishings	**1.34**			
2 Furnishings for routine care	Space and Furnishings	1.24			
3 Furnishings for relaxation	Space and Furnishings	.97	1.18	1.09	.99
4 Room Arrangement	Space and Furnishings	1.19			
5 Space for Privacy	Space and Furnishings	.88	1.05	1.08	1.18
6 Display for Children	Space and Furnishings	.92			
7 Gross Motor Space	Space and Furnishings	**1.39**			
8 Gross Motor Equipment	Space and Furnishings	**1.59**			
9 Greeting/Departing	Personal Care Routines	**1.75**			
10 Meals/Snacks	Personal Care Routines	**1.82**			
11 Nap	Personal Care Routines	1.29			

Table 7. *Rasch Outfit Statistics for Each Specification of the Measure*

Items	Subscale	$Outfit^{37}$	$Outfit^{16}$	$Outfit^{PT}$	$Outfit^{SP}$
12 Diapering/Toileting	Personal Care Routines	**1.95**			
13 Health Practice	Personal Care Routines	**1.46**			
14 Safety Practice	Personal Care Routines	**1.96**			
15 Books and Pictures	Language and Reasoning	.74	.88	.9	1.12
16 Encouraging to Communicate	Language and Reasoning	.81			
17 Language to develop reasoning	Language and Reasoning	.98	1.2	1.2	.98
18 Informal use of language	Language and Reasoning	.93	.99	.92	.79
19 Fine motor activities	Activities	.79	.86	.81	.9
20 Art	Activities	.7	.77	.77	.88
21 Music and Movement	Activities	.78			
22 Blocks	Activities	.71	.84	.74	.7
23 Sand and water play	Activities	1.09			
24 Dramatic Play	Activities	**.67**	.78	.7	.6
25 Nature and Science	Activities	.84	.98	.98	.98
26 Math	Activities	.71	.78	.79	.96
27 Use of TV, video or computer	Activities	1.21			
28 Promoting acceptance of diversity	Activities	.94			
29 Gross motor supervision	Interaction	.98			
30 General supervision	Interaction	1.05	**1.35**	1.21	1.09
31 Discipline	Interaction	.81	.9	.96	1.12
32 Staff-child interactions	Interaction	**1.44**	**1.5**	1.17	1.12
33 Interactions among children	Interaction	.88	.89	.9	.77
34 Schedule of daily play	Program Structure	1.21			
35 Free play	Program Structure	.97			
36 Group time	Program Structure	1.09	1.2	**1.72**	**1.32**
37 Provisions for exceptional children	Program Structure	**1.62**			

Note. Values in bold signify values outside of the recommended fit criteria. 37 = Unidimensional Specification of the 37-Item; 16 = 16 Item Unidimensional Specification; PT = Cassidy et. al. (2005^{a}) 2-Dimensional Specification; SP = Cassidy et. al. (2005^{b}) 2-Dimension Within Item Dimensional Specification. The T statistics are not presented because of the sensitivity of these statistics to large sample sizes

Table 8. *Rasch Infit Statistics for Each Specification of the Measure*

Items	Subscale	$Infit^{37}$	$Infit^{16}$	$Infit^{PT}$	$Infit^{SP}$
1 Indoor Space	Space and Furnishings	1.24			
2 Furnishings for routine care	Space and Furnishings	1.23			

Table 8. *Rasch Infit Statistics for Each Specification of the Measure*

Items	Subscale	$Infit^{37}$	$Infit^{16}$	$Infit^{PT}$	$Infit^{SP}$
3 Furnishings for relaxation	Space and Furnishings	.93	1.17	1.11	.99
4 Room Arrangement	Space and Furnishings	1.18			
5 Space for Privacy	Space and Furnishings	.86	1.06	1.11	1.13
6 Display for Children	Space and Furnishings	.83			
7 Gross Motor Space	Space and Furnishings	**1.34**			
8 Gross Motor Equipment	Space and Furnishings	**1.54**			
9 Greeting/Departing	Personal Care Routines	**1.57**			
10 Meals/Snacks	Personal Care Routines	**1.92**			
11 Nap	Personal Care Routines	1.27			
12 Diapering/Toileting	Personal Care Routines	**1.94**			
13 Health Practice	Personal Care Routines	**1.53**			
14 Safety Practice	Personal Care Routines	**2**			
15 Books and Pictures	Language and Reasoning	**.67**	.85	.89	.98
16 Encouraging to Communicate	Language and Reasoning	.81			
17 Language to develop reasoning	Language and Reasoning	.98	1.22	1.2	1.01
18 Informal use of language	Language and Reasoning	.92	1.01	.97	.85
19 Fine motor activities	Activities	.76	.86	.83	.91
20 Art	Activities	**.68**	.77	.79	.84
21 Music and Movement	Activities	.77			
22 Blocks	Activities	**.67**	.82	.74	.7
23 Sand and water play	Activities	1.04			
24 Dramatic Play	Activities	**.6**	.72	**.68**	.6
25 Nature and Science	Activities	.86	1.02	1.03	1
26 Math	Activities	**.67**	.77	.78	.87
27 Use of TV, video or computer	Activities	1.22			
28 Promoting acceptance of diversity	Activities	.87			
29 Gross motor supervision	Interaction	.95			
30 General supervision	Interaction	1.1	**1.37**	1.3	1.18
31 Discipline	Interaction	.83	.91	.84	1.04
32 Staff-child interactions	Interaction	**1.58**	**1.71**	1.51	1.42
33 Interactions among children	Interaction	.97	1.02	.95	.88
34 Schedule of daily play	Program Structure	1.21			
35 Free play	Program Structure	.96			
36 Group time	Program Structure	1.18	1.29	**1.8**	**1.32**
37 Provisions for exceptional children	Program Structure	**1.62**			

Table 8. *Rasch Infit Statistics for Each Specification of the Measure*

Items	Subscale	$Infit^{37}$	$Infit^{16}$	$Infit^{PT}$	$Infit^{SP}$

Note. Values in bold signify values outside of the recommended fit criteria. 37 = Unidimensional Specification of the 37-Item; 16 = 16 Item Unidimensional Specification; PT = Cassidy et. al. (2005^{a}) 2-Dimensional Specification; SP = Cassidy et. al. (2005^{b}) 2-Dimension Within Item Dimensional Specification. The T statistics are not presented because of the sensitivity of these statistics to large sample sizes

Item Technical Quality for the 37-Item Unidimensional Specification. Nine items from this specification of the measure displayed misfit greater than the threshold of 1.3, with: "Indoor Space" (Outfit = 1.34), "Gross Motor Development" (Outfit = 1.39, Infit = 1.34), "Gross Motor Equipment" (Outfit = 1.59, Infit = 1.54), "Greeting/Departing" (Outfit = 1.75, Infit = 1.57), "Meals/Snacks" (Outfit = 1.82, Infit = 1.92), "Diapering/Toileting" (Outfit = 1.95, Infit = 1.94), "Health and Practice" (Outfit = 1.46, Infit = 1.53), "Safety Practice" (Outfit = 1.96, Infit = 2.00),"Staff-child Interactions" (Outfit = 1.44, Infit = 1.58), and "Provisions for Exceptional Children" (Outfit = 1.62, Infit = 1.62). These nine items exhibited deficiencies in the model predicted randomness, and as such were more unpredictable than the model would have predicted. Five additional items from this specification of the measure displayed misfit that was below the .7 threshold, with "Books and Pictures" (Infit = .67), "Art" (Infit = .68), "Blocks" (Infit = .67), "Dramatic Play" (Outfit = .67, Infit = .60), and "Math" (Infit = .67). The misfit of these items indicated that they were too predictable for the model.

Next, the patterns of misfitting items in the measure were examined for issues in the presentation order of items. The presentation order of items often refers to the extent to which responses to one item influence responses to other items in the measure (Marais & Andrich, 2008). The influence of the presentation order of items is revealed through the concentration of item misfit among items that occur in close proximity in the measure (Meijer, 2003). The patterns of misfit in this specification of the measure exhibited possible issues with the presentation order of items, as two locations in the measure contained concentrations of items

that exhibited misfit. These regions of misfit included: (1) the "Gross Motor Development", "Gross Motor Equipment", "Greeting/Departing", and the "Meals/Snacks", which occurred at roughly the same location in the scale (i.e., items 7,8,9 and 10, respectively), and (2) the "Diapering/Toileting", "Health and Practice", and "Safety Practice", which were the 12th, 13th, and 14th items within the measure.

Item Technical Quality for the 16-Item Unidimensional Specification. Two items from this specification of the measure exhibited misfit, with the “General Supervision” item showing (Infit = 1.37 and Outfit = 1.35), and the “Staff-child Interactions” items showing (Infit = 1.71 and Outfit = 1.5). Both items were too unpredictable for the model. The measure showed no discernable issues with the presentation order of items.

Item Technical Quality for the Provisions for Learning/Teaching Interactions 2-Dimensional Specification. The item fit statistics for this specification of the measure showed that the “Group Time” item was too unpredictable for the model, with Infit = 1.72 and Outfit = 1.8. Further, the “Dramatic Play” item was too predictable for the model, with Infit = .68. No notable patterns of misfit were observed in relation to item presentation ordering.

Item Technical Quality for the Structural/Process 2-Dimensional Within-Item Dimensional Specification. Only one item from this specification of the measure exhibited marginal misfit. Specifically, the "Group time" item exhibited underfit to the model, with Outfit = 1.32 and Infit = 1.32.

Rating Scale Functioning. The goal of these analyses was to establish what adjustments to the rating scale were needed to establish monotonic functioning, and to establish whether the hypothesis about the sufficiency of a 4-point rating scale was supported. The monotonic functioning of the rating scale was investigated through examining the estimated average

measures for each level of the rating scale. Specifically, it was critical to investigate that higher categories were associated with higher classroom quality scores (Linacre, 1999; 2002). The average estimates for each category of the rating scale are reported in Table 9. Results showed that the rating scale was disordered across every specification of the measure. Specifically, both categories 4 and 6 of the rating scale exhibited disordered categories. In other words, each level of the rating scale was not substantively associated with a higher level of functioning on the underlying continuum of classroom quality. As a result, using the 7-item rating scale for this measure was considered faulty, and revisions to the rating scale were sought (Linacre, 2010).

Table 9. *Average Difficulty for Each Level of the Rating Scale for Each Model*

Category	Average Measure	SE
Unidimensional 37-Item Specification		
1	-5.63	.01
2	.27	.01
3	1.23	.01
4	**-.05**	.01
5	2.71	.01
6	**.63**	.01
7	NA	NA
Unidimensional 16-Item Specification		
1	-5.86	.02
2	.33	.02
3	.43	.02
4	**-.02**	.01
5	3.13	.02
6	**.88**	.02
7	NA	NA
Provisions for Learning/Teaching and Interactions Specification		
1	-6.5	.02
2	.25	.02
3	.41	.02
4	**.01**	.02
5	3.23	.02
6	**1.08**	.02
7	NA	NA
Structural/Process Within-item Dimensional Specification		
1	-2.97	.02
2	-.32	.02

Table 9. *Average Difficulty for Each Level of the Rating Scale for Each Model*

Category	Average Measure	SE
3	-.06	.02
4	**-.50**	.01
5	2.66	.02
6	**.47**	.02
7	NA	NA

Note. Values in bold signify categories that exhibited disordered categories. Category 7 is anchored for estimation in the Andrich Rating Scale Model. Also, it is necessary to anchor one category of the rating scale (i.e., listed in the table above as category 7, which has an estimate of "NA"). As a result, an estimated for the level of the rating scale that is anchored cannot be provided in the analysis.

In order to establish a monotonically increasing rating scale, category 2 was combined with category 3, category 4 was combined with category 5, and category 6 was combined with category 7. These adjustments to the rating scale were in line with what was observed in Table 2, whereby these categories were sparsely utilized by raters. Further, these revisions to the rating scale were also in line with how the rating scale for each item is explained in measure, as only the odd numbers in the rating scale of the ECERS-R (i.e., 1, 3 5, and 7) offered indicators for quality. The estimates for the average measures associated with each increase in the categories for the revised rating scale are found in Table 10. This revised rating scale exhibited monotonic functioning, which was evidenced by the average measure score increasing for each increase in the category rating scale, and confirmed the hypothesis that a 4-point rating scale was appropriate. Consequently, all subsequent analyses utilized this 4-point rating scale.

Table 10. *Average Difficulty for Each Level of the Rating Scale for Each Model With Collapsed Categories in Rating Scale*

Category	Average Measure	SE
Unidimensional 37-Item Specification		
1	-5.47	.01
2	.75	.01
3	1.93	.01
4	NA	NA
Unidimensional 16-Item Specification		
1	-5.92	.02
2	.52	.02
3	1.81	.02

Table 10. *Average Difficulty for Each Level of the Rating Scale for Each Model With Collapsed Categories in Rating Scale*

Category	Average Measure	SE
4	NA	NA
Provisions for Learning/Teaching and Interactions Specification		
1	-6.30	.02
2	.42	.02
3	1.84	.02
4	NA	NA
Structural/Process Within-item Dimensional Specification		
1	-7.20	.02
2	.78	.02
3	2.26	.02
4	NA	NA

Note. Values in bold are disordered categories. Category 4 is anchored for estimation in the Andrich Rating Scale Model.

Revised Item Technical Quality. As a consequence of revising the rating scale, the item fit statistics were reexamined. Information pertaining to the revised technical quality of the items are reported in Tables 11 and 12.

Table 11. *Rasch Outfit Statistics for Specifications with Collapsed Categories*

Items	Subscale	Outfit^{37}	Outfit^{16}	Outfit^{PT}	Outfit^{SP}
1 Indoor Space	Space and Furnishings	1.21			
2 Furnishings for routine care	Space and Furnishings	1.17			
3 Furnishings for relaxation	Space and Furnishings	.94	1.20	1.13	1.04
4 Room Arrangement	Space and Furnishings	1.06			
5 Space for Privacy	Space and Furnishings	.87	1.05	1.09	1.26
6 Display for Children	Space and Furnishings	.93			
7 Gross Motor Space	Space and Furnishings	1.25			
8 Gross Motor Equipment	Space and Furnishings	1.44			
9 Greeting/Departing	Personal Care Routines	1.51			
10 Meals/Snacks	Personal Care Routines	1.58			
11 Nap	Personal Care Routines	1.08			
12 Diapering/Toileting	Personal Care Routines	1.61			
13 Health Practice	Personal Care Routines	1.08			
14 Safety Practice	Personal Care Routines	**1.69**			
15 Books and Pictures	Language and Reasoning	**.6**	.79	.79	1.18
16 Encouraging to Communicate	Language and Reasoning	.77			

Table 11. *Rasch Outfit Statistics for Specifications with Collapsed Categories*

Items	Subscale	Outfit^{37}	Outfit^{16}	Outfit^{PT}	Outfit^{SP}
17 Language to develop reasoning	Language and Reasoning	.93	1.17	1.31	1.04
18 Informal use of language	Language and Reasoning	.79	.91	.87	.79
19 Fine motor activities	Activities	**.65**	.79	.76	.86
20 Art	Activities	**.66**	.79	.83	1
21 Music and Movement	Activities	.7			
22 Blocks	Activities	**.71**	.90	.82	.78
23 Sand and water play	Activities	1.1			
24 Dramatic Play	Activities	**.68**	.88	.79	**.69**
25 Nature and Science	Activities	.74	.90	.89	1.04
26 Math	Activities	**.67**	.79	.8	1.02
27 Use of TV, video or computer	Activities	1.07			
28 Promoting acceptance of diversity	Activities	.94			
29 Gross motor supervision	Interaction	.99			
30 General supervision	Interaction	.99	**1.34**	1.21	1.13
31 Discipline	Interaction	.88	1.01	.92	1.21
32 Staff-child interactions	Interaction	1.19	**1.32**	1.03	1.03
33 Interactions among children	Interaction	.9	.96	.79	.84
34 Schedule of daily play	Program Structure	.95			
35 Free play	Program Structure	.84			
36 Group time	Program Structure	1.08	1.27	**1.73**	1.34
37 Provisions for exceptional children	Program Structure	**1.6**			

Note. Values in bold signify items outside of the recommended fit criteria. 37 = Unidimensional Specification of the 37-Item; 16 = 16 Item Unidimensional Specification; PT = Cassidy et. al. (2005[a]) 2-Dimensional Specification; SP = Cassidy et. al. (2005[b]) 2-Dimension Within Item Dimensional Specification. The T statistics are not presented because of the sensitivity of these statistics to large sample sizes

Table 12. *Rasch Infit Statistics for Specifications with Collapsed Categories*

Items	Subscale	Infit^{37}	Infit^{16}	Infit^{PT}	Infit^{SP}
1 Indoor Space	Space and Furnishings	1.16			
2 Furnishings for routine care	Space and Furnishings	**1.46**			
3 Furnishings for relaxation	Space and Furnishings	.93	1.19	1.13	1.05
4 Room Arrangement	Space and Furnishings	1.1			
5 Space for Privacy	Space and Furnishings	.87	1.06	1.1	1.19
6 Display for Children	Space and Furnishings	.86			
7 Gross Motor Space	Space and Furnishings	1.18			
8 Gross Motor Equipment	Space and Furnishings	1.44			

Table 12. *Rasch Infit Statistics for Specifications with Collapsed Categories*

Items	Subscale	$Infit^{37}$	$Infit^{16}$	$Infit^{PT}$	$Infit^{SP}$
9 Greeting/Departing	Personal Care Routines	1.48			
10 Meals/Snacks	Personal Care Routines	1.65			
11 Nap	Personal Care Routines	1.04			
12 Diapering/Toileting	Personal Care Routines	1.63			
13 Health Practice	Personal Care Routines	1.08			
14 Safety Practice	Personal Care Routines	**1.7**			
15 Books and Pictures	Language and Reasoning	**.57**	.74	.77	.91
16 Encouraging to Communicate	Language and Reasoning	.86			
17 Language to develop reasoning	Language and Reasoning	.94	1.17	1.28	1.07
18 Informal use of language	Language and Reasoning	.83	.97	.94	.86
19 Fine motor activities	Activities	**.65**	.79	.78	.88
20 Art	Activities	**.65**	.78	.82	.9
21 Music and Movement	Activities	**.65**			
22 Blocks	Activities	.7	.89	.81	.79
23 Sand and water play	Activities	1.12			
24 Dramatic Play	Activities	**.63**	.81	.76	**.68**
25 Nature and Science	Activities	.75	.90	.89	.99
26 Math	Activities	**.66**	.78	.79	.92
27 Use of TV, video or computer	Activities	1.07			
28 Promoting acceptance of diversity	Activities	.89			
29 Gross motor supervision	Interaction	1			
30 General supervision	Interaction	1.13	**1.46**	**1.36**	1.29
31 Discipline	Interaction	.95	1.10	.93	1.23
32 Staff-child interactions	Interaction	1.49	**1.72**	**1.47**	**1.47**
33 Interactions among children	Interaction	1.13	1.27	1.08	1.12
34 Schedule of daily play	Program Structure	.91			
35 Free play	Program Structure	.87			
36 Group time	Program Structure	1.26	**1.48**	**1.96**	**1.52**
37 Provisions for exceptional children	Program Structure	**1.6**			

Note. Values in bold signify categories that exhibited disordered categories. 37 = Unidimensional Specification of the 37-Item; 16 = 16 Item Unidimensional Specification; PT = Cassidy et. al. (2005[a]) 2-Dimensional Specification; SP = Cassidy et. al. (2005[b]) 2-Dimension Within Item Dimensional Specification. The T statistics are not presented because of the sensitivity of these statistics to large sample sizes

Revised Item Technical Quality for the 37-Item Unidimensional Specification. The following items for this specification had fit statistics below the .7 threshold: “Books and Pictures” (Outfit = .60 and Infit = .57), “Fine Motor Activities” (Outfit = .65 and Infit = .65), “Art” (Outfit = .66 and Infit = .65), “Dramatic Play” (Outfit = .68 and Infit = .63), “Music and Movement” (Infit = .65), “Math” (Outfit = .67 and Infit = .66), and "Blocks" (Outfit = .71). However, the decision was made to retain these items in subsequent analyses because of the fact that the items overfit the model. Three of these items showed a potential issue with presentation order, "Fine Motor Activities", "Art," and "Music and Movement,” as each occurred at approximately the same location in the measure. Additionally, three items contained item fit statistics greater than the 1.3 threshold, which indicated that these items were too unpredictable for the model. Specifically, "Safety Practice" (Outfit = 1.69, Infit = 1.70), "Provisions for exceptional children" (Outfit = 1.6, Infit = 1.6), and "Furnishings for routine care" (Infit = 1.46). Only two of these items exceeded the threshold of 1.3 on both the Infit and Outfit Statistic. However, given the variation in reported criteria for misfit, and novel nature of this analysis, the item was retained for parity in the subsequent analyses.

Revised Item Technical Quality for the 16-Item Unidimensional Specification. Three of the items in this specification of the measure showed some kind of misfit. Specifically, the “General Supervision” (Outfit = 1.34, Infit = 1.46) and “Staff Child Interactions” (Outfit = 1.32, Infit = 1.72), and “Group Time” (Infit = 1.48) items. All of these items showed misfit above the threshold of a mean-squared value of 1.3, which indicated that responses to these items were more unpredictable than would be expected. However, because these items were on the margin of the recommended threshold, the items were retained in subsequent analyses.

Revised Item Technical Quality for the Provisions for Learning/Teaching Interactions 2-Dimensional Specification. Three items from this specification of the measure exhibited misfit. The "Group Time" showed misfit on both fit statistics, with Outfit = 1.73 and Infit = 1.96, whereas the "General Supervision" (Infit = 1.36) and "Staff Child Interactions" (Infit = 1.47) exhibited misfit for just the Infit statistic. Because the outfit statistic for the "Group Time" item only marginally misfit, and because the "General Supervision" and "Staff Child Interactions" items only misfit on one fit statistic, these items were retained in subsequent analyses.

Revised Item Technical Quality for the Structural/Process 2-Dimensional Within-Item Dimensional Specification. Only two items from this specification of the measure marginally underfit the model, with "Staff-child Interactions" (Infit = 1.47) and "Group Time (Outfit = 1.34, Infit = 1.52). However, only the "Group Time" item contained items that exhibited misfit for both the Infit and Outfit statistics. Further, the "Dramatic Play" item overfit the model, with Outfit = .68, but did not show misfit for the Infit statistic. Again, given the novel nature of this work, and the different criteria for item misfit, these items were retained in the subsequent analyses.

Summary of Revised Item Technical Quality. Results showed that the proportion of items that exhibited misfit, all of the 16-item specifications of the measure were superior to the 37-item specification of the measure. Further, the two multidimensional specifications of the measure contained few items that misfit the underlying Rasch model, with the Structural/Process specification containing the fewest items that misfit its respective Rasch model.

Expected Item Difficulty Hierarchy. Two analyses were carried out in order to provide evidence for the hypothesis that the interactional items from the measure would be more difficult to endorse than items that addressed more of the structural features in the classroom. First,

descriptive statistics for each dimension of the multidimensional specifications were examined to see if they conformed to theoretical expectations. Second, the hierarchy of item difficulty calibrations was examined across all specifications of the measure.

In order to provide evidence to support the substantive aspect of Messick's (1989) validity framework, the average of the classroom Rasch measure scores were examined across each model specification. The means and standard deviations for these measure scores are provided in Table 13 below. Dimensions for each model specification are ordered from most difficult to least difficult, with the classroom quality Rasch estimates for the two unidimensional specifications provided for context.

Table 13. *Dimensional Ordering of Mean Population Parameters in Logits*

Unidimensional 37		Unidimensional 16		Provisions for Learning/Teaching Interactions 2-Dimension			Structural/Process 2-Dimension Within-Item		
Mean (SE)	SD	Mean (SE)	SD		Mean (SE)	SD		Mean (SE)	SD
.29	.90	1.37	1.29	Teaching Interactions	.02	1.81	Structural	.73	1.26
				Provisions for Learning	1.15	1.38	Process	1.26	1.45

Note. SD = Standard Deviation; SE = Standard Error.

The dimensional ordering demonstrated that, for the Provisions for Learning/Teaching Interactions 2-dimension specification of the measure, items from the Teaching and Interactions Dimension (Mean = .02; *SD* =1.81) were more difficult to endorse than the Provisions for Learning Dimension (Mean = 1.15; *SD* =1.38). The dimensional ordering of the Structural/Process 2-dimension within-item specification showed that the items from the Structural Dimension were on average more difficult to endorse (Mean = .73; *SD* = 1.26) than those from the Process Dimension (Mean = 1.26; *SD* = 1.45). Results for the Provisions for

Learning/Teaching Interactions specification of the measure were in line with what was hypothesized, while the results for the Structural/Process diverged from what was expected.

The stability of the item difficulty estimates was examined across all model specifications. The item difficulty calibrations for each dimension are found in Table 14 below, and have been sorted from most difficult to least difficult (i.e, ascending order of difficulty). The 37-item unidimensional model specification is provided for context, but some care should be taken when examining the tables, as the ascending order of items in this specification of the measure does not match the other model specifications due to the different number of items in the measure. The item hierarchy was identical across the 16-item unidimensional model, the Provisions for Learning/Teaching and Interactions specification, and the Structural/Process specification. Further, the item hierarchy from these specifications was more generally found in the 37-item unidimensional specification the measure, with no disorder observed in the ordering of the item difficulty of the 16-item versions of the measure found in the 37-item unidimensional specification. These results indicated that the hierarchy of item difficulty estimates was invariant across different specifications of the measure. However, what was notable across all of these specifications was that the items pertaining to interactions between children and teachers were the easiest to endorse.

Table 14. *Comparing the Rank Order of Item Difficulty Estimates Across all Model Specifications*

37-item Unidimensional Specification		16-item Unidimensional Specification		Provisions for Learning/Teaching Interactions 2-Dimension		Structural/Process 2-Dimension Within-Item	
Item	d	Item	d	Item	d	Item	d
10 Meals/Snacks	-1.16	25 Nature and Science	-1.50	25 Nature and Science	-.51	25 Nature and Science	-0.82
12 Diapering/Toileting	-1.43	20 Art	-2.02	20 Art	-1.19	20 Art	-1.58
25 Nature and Science	-1.50	24 Dramatic Play	-2.02	24 Dramatic Play	-1.19	24 Dramatic Play	-1.58
11 Nap	-1.51	26 Math	-2.05	26 Math	-1.23	26 Math	-1.62
14 Safety Practice	-1.66	5 Space for Privacy	-2.07	5 Space for Privacy	-1.25	5 Space for Privacy	-1.65
27 Use of TV, video or computer	-1.68	3 Furnishings for relaxation	-2.13	3 Furnishings for relaxation	-1.34	3 Furnishings for relaxation	-1.74
21 Music and Movement	-1.72	22 Blocks	-2.21	22 Blocks	-1.44	22 Blocks	-1.86

Table 14. *Comparing the Rank Order of Item Difficulty Estimates Across all Model Specifications*

37-item Unidimensional Specification		16-item Unidimensional Specification		Provisions for Learning/Teaching Interactions 2-Dimension		Structural/Process 2-Dimension Within-Item	
Item	d	Item	d	Item	d	Item	d
13 Health Practice	-1.75	17 Language to develop reasoning	-2.26	17 Language to develop reasoning	-1.51	17 Language to develop reasoning	-2.04
7 Gross Motor Space	-1.91	15 Books and Pictures	-2.33	15 Books and Pictures	-1.60	15 Books and Pictures	-2.42
23 Sand and water play	-1.95	19 Fine motor activities	-2.60	19 Fine motor activities	-1.95	19 Fine motor activities	-2.94
28 Promoting acceptance of diversity	-1.97	30 General supervision	-3.02	30 General supervision	-2.50	30 General supervision	-4.23
20 Art	-2.02	18 Informal use of language	-3.08	18 Informal use of language	-2.57	18 Informal use of language	-4.32
24 Dramatic Play	-2.02	31 Discipline	-3.08	31 Discipline	-2.58	31 Discipline	-4.34
26 Math	-2.05	36 Group time	-3.14	36 Group time	-2.65	36 Group time	-4.43
6 Display for Children	-2.06	33 Interactions among children	-3.57	33 Interactions among children	-3.2	33 Interactions among children	-5.14
5 Space for Privacy	-2.07	32 Staff-child interactions	-3.64	32 Staff-child interactions	-3.29	32 Staff-child interactions	-5.25
3 Furnishings for relaxation	-2.13						
22 Blocks	-2.21						
34 Schedule of daily play	-2.24						
17 Language to develop reasoning	-2.26						
8 Gross Motor Equipment	-2.33						
15 Books and Pictures	-2.33						
29 Gross motor supervision	-2.56						
19 Fine motor activities	-2.6						
35 Free play	-2.77						
1 Indoor Space	-2.88						
4 Room Arrangement	-2.89						
30 General supervision	-3.02						
18 Informal use of language	-3.08						
31 Discipline	-3.08						
36 Group time	-3.14						
37 Provisions for exceptional children	-3.28						
16 Encouraging to Communicate	-3.39						
9 Greeting/Departing	-3.50						
33 Interactions among children	-3.57						
32 Staff-child interactions	-3.64						
2 Furnishings for routine care	-4.40						

Note. d = item difficulty estimate.

Generalizability Validity Evidence

Next, analyses were carried out in order to examine whether the observed measures scores were generalizable to the population. The primary aims of these analyses were to examine the stability of scores across different model specifications and to investigate invariance of items

across subgroups. Results are presented for the following: (1) stability of items across model specifications; (2) comparing Rasch reliability statistics and person estimates; (3) precision of the person estimates across models, (4) tests for differential item functioning (i.e., item bias). These analyses were conducted to answer research questions pertaining to the reliability of the measured dimensions, as well as for the questions pertaining to possible item bias in the measure.

Item Difficulty Invariance. Pearson correlation coefficients were utilized to examine associations between pairs of item difficulties for each model specification. The goal with these analyses was to observe the stability of the item difficulty estimates for the items across all model specifications. The correlations are reported in Table 15 below. The item difficulties exhibited almost perfect correlations across the two unidimensional specifications of the measure, as well as for the Provisions for Learning/Teaching Interactions 2-dimension specification. The Structural/Process specification was also positive and strongly associated with the other model specifications. As a result, the item difficulty estimates were considered invariant across all model specifications.

Table 15. *Item Correlations Between Pairs of Item Difficulty Estimates*

	1	2	3	4
Unidimensional[37]	1			
Unidimensional[16]	.99	1		
2-Dimension[PT]	.98	.98	1	
2-Dimension Within[SP]	.90	.90	.85	1

Note. 37 = 37-item unidimensional specification and 16 = 16-item unidimensional specification. The correlations with the 37-item unidimensional scores are constrained to only the 16 items used in the other model specifications. PT = Provisions for Learning/Teaching and Interactions. 2-Dimension specification. SP = Structure/Process 2-Dimension Within-Item Dimensional Specification.

Reliability of Person Estimates. Rasch reliability statistics were examined to investigate the extent to which items distinguished between distinct levels of classroom quality. The

expected *a posteriori* (EAP) reliability coefficients[9] are reported in Table 16 below. The reliability coefficients across all model specifications were > .70, which indicated that the internal consistency of the items across all model specifications was good. Both unidimensional specifications of the measure had the highest EAP reliabilities, with EAP = .93 for the 37-item unidimensional specification, and EAP = .90 for the 16-item unidimensional specification. Both dimensions of the Structural/Process 2-dimension within-item specification of the measure showed the lowest EAP reliabilities, with EAP = .75 for the Structural Dimension, and EAP = .82 for the Process Dimension. The reliabilities for the Provisions for Learning and Teaching and Interactions Dimensions from the Provisions for Learning/Teaching Interactions 2-dimension specification where EAP = .87 and EAP = .86, respectively.

In order to provide context about the differences in reliability estimates between dimensions, the Spearman Brown's formula[10] was used to calculate the number of additional items that would be necessary for each dimension from the multidimensional specifications of the measure to reach the EAP reliabilities of the two unidimensional specifications of the measure. These figures are also reported in Table 16. Analyses showed that both the Provisions for Learning and Teaching and Interactions dimensions would need one additional item to reach the EAP reliability of the 16-item unidimensional specification, and two additional items to reach the EAP reliability of the 37-item unidimensional specification of the measure. The Structural dimension from the Structural/Process 2-dimension within-item specification of the measure needed four items to reach the EAP reliability of the 37-item unidimensional specification of the

[9] The reliability formulas follow Adams (2005): EAP reliability is defined as 1 - s/(s+v) = v/(s+v), where v denotes the variance of theta estimates and s denotes the average of the squared error.

[10] $N = \frac{\rho^{*}_{xx\prime}(1-\rho_{xx\prime})}{\rho_{xx\prime}(1-\rho^{*}_{xx\prime})}$, where N = the additional number of items needed to obtain a given reliability, $\rho_{xx\prime}$ is the desired benchmark reliability (i.e., either unidimensional specification of the measure), and $\rho^{*}_{xx\prime}$ is reliability of the current measure.

measure, and three items to reach the EAP reliability estimate for the 16-item unidimensional specification of the measure. Finally, the Process dimension would need two and three additional items to match the EAP estimates from each of the respective unidimensional specifications of the measure. These results suggest that in choosing between the two multidimensional specifications of the measure, the Provisions for Learning/Teaching Interactions specification (i.e., containing Provisions for Learning and Teaching and Interactions) exhibited better reliability than the Structural/Process specification (i.e., containing Structural and Process dimensions). However, both multidimensional specifications exhibited excellent reliability given the small number of items contained within each dimension (Cortina, 1993; Green, Lissityz & Mulaik, 1977).

Table 16. *Rasch EAP Reliability and Spearman Brown Additional Items*

Model	# of Items	EAP	Spearman Brown Additional Items (Ref. 37-item Unidimensional Measure)	Spearman Brown Additional Items (Ref. 16-item Unidimensional Measure)
Unidimensional	37	.93	-	-
Unidimensional	16	.90	1.48	-
Provisions for Learning[a]	8	.87	1.99	1.34
Teaching and Interactions[a]	8	.86	2.16	1.47
Structural[b]	11	.75	4.43	3.00
Process[b]	12	.82	2.92	1.98

Note. a = a dimension from the Cassidy (2005[a]) 2-dimension specification of the measure. b = items from the Structural/Process 2-dimension within-item specification of the measure. G = Person Separation Index, H = Person Strata Index.

Differential Item Functioning. Differential item tests were used to provide information for the hypotheses that items were likely biased in favor of Head Start centers and teachers with Graduate degrees. Estimates for items which exhibited notable DIF are found in Table 17, while Appendix A provides the DIF estimates for all combinations of covariates and items.

The 37-item unidimensional specification of the measure exhibited DIF for 24.32% of the items. The DIF results were as follows: the "Meals and Snacks" item was easier for Head Start programs (DIF = -.64); the "Greeting/Departing" (DIF = -1.18), "Informal Use of Language" (DIF = -.62), "Discipline" (DIF = -.85), and "Free Play" (DIF = -.82) items were all easier for private centers; "Free Play" (DIF =.54) was harder for state and local programs. The following items were easier for teachers with only a vocational degree: "Space for Privacy" (DIF = -.72), "Gross Motor Space" (DIF = -.57), and "Meals/Snacks" (DIF = -.53). For teachers with a graduate degree, the "Gross Motor Space" and "Gross Motor Equipment" items were more difficult, while the "Encouraging Child Interactions" was easier. The 16-item unidimensional specification of the measure had only one item (i.e., 6.25% of the measure) that exhibited item bias. The "Art" item was easier for teachers with a vocational degree (DIF = -.51), and was more difficult for teachers with a graduate degree, with (DIF = .71). The Provisions for Learning/Teaching Interactions 2-dimension specification of the measure showed DIF for three items (i.e., 18.75% of the measure). The "Fine Motor Activities" item was easier for Head Start centers (DIF = -.64), and more difficult for private programs (DIF = .92). In addition, the "Discipline" (DIF = -.50) item was easier for private centers, and the "Staff Child Interactions" item was harder (DIF = .60). Finally, the Structural/Process 2-dimension within-item specification of the measure showed no DIF for any of the demographic variables examined.

Summary of Item Bias. Overall, while evidence for some item bias exits across all specifications of the measure, it is most problematic in the 37-item unidimensional specification of the measure. Most of the DIF for this specification of the measure was related to center types and teacher educational status. No discernable patterns were of note. A similar theme was observed for both the 16-item unidimensional specification of the measure and the Provisions for

Learning/Teaching Interactions 2-dimension specification. Most notable was that the Structural/Process showed no item bias. However, given the number of possible DIF contrasts tested, all specifications of the measure exhibited very little item bias.

Table 17. *Notable Differential Item Functioning for All Model Specifications*

Item	Category	DIF Est	SE	Est/SE
Unidimensional 37-item Specification				
Meals and Snacks	Head Start	-.64	.06	-10.83
Greeting/Departing	Private	-1.18	.08	-14.75
Informal Use of Language	Private	-.62	.07	-8.86
Discipline	Private	-.85	.07	-12.14
Free Play	Private	-.82	.07	-11.71
Free Play	State & Local	.54	.05	10.80
Furnishings for Routine Care	Less Than High School	1.38	.23	6
Space for Privacy	Vocational	-.72	.23	-3.13
Gross Motor Space	Vocational	-.57	.22	-2.59
Meals/Snacks	Vocational	-.53	.22	-2.40
Gross Motor Space	Graduate	.56	.07	8
Gross Motor Equipment	Graduate	.92	.07	13.14
Encouraging to Communicate	Graduate	-.51	.09	-5.67
Unidimensional 16-item Specification				
Art	Vocational	-.51	.21	-2.43
Art	Graduate	.71	.07	10.14
Provisions for Learning/Teaching and Interactions Specification				
Fine Motor Activities	Head Start	-.64	.06	-10.67
Fine Motor Activities	Private	.92	.06	15.33
Discipline	Private	-.50	.06	-8.33
Staff Child Interactions	Private	.60	.06	10

Precision of Person Estimates. The estimated classroom quality measures (i.e., "Rasch scores") were plotted against their standard errors in order to examine the precision[11] of the classroom quality measures across the underlying continua for classroom quality (Wolfe &

[11] To understand the precision of the measure it is useful to draw a parallel to an archer shooting arrows at a target: Precision is analogous to the clustering together of arrows shot at a target, where by increased levels of precision are related to a tighter clustering of arrows which have been shot at a target.

Smith, 2007). The goals with these analyses were to compare the precision of measures across the different specifications of the measure, and to look within each specification of the measure to identify where each measure was most precise along the underlying latent continua for quality. A graphical display of the estimated classroom quality Rasch measures plotted against their associated standard errors is found in Figure 3. Each specification of the measure has been plotted and color coded for direct comparison of the precision across the different specifications of the measure.

Figure 3. Precision of the Classroom Measure Estimates. A random sample of 150 classrooms are plotted to ease the interpretation of the data points.

In interpreting Figure 3, it was useful to examine where scores were most precise along the latent continuum. This was established by examining where the approximate inflection point

occurred within each curve, and comparing that point across the different specifications of the measure. The 37-item unidimensional specification of the measure appeared to be more precise around -.75 logits. The general shape of the curve indicated that classrooms with lower levels of quality were measured more accurately than classrooms with higher levels of quality. The 16-item unidimensional specification of the measure appeared to hit an inflection point of around -1.25 logits, and also measured classrooms with lower quality more precisely than classrooms with higher quality. The Provisions for Learning dimension appeared to hit an inflection point of about 0 logits. The Teaching and Interactions dimension hit an inflection point at about 0 logits, with the shape of the points again indicating classrooms with lower levels of quality were measured with more precision. Both the Structural and Process dimensions showed a similar pattern, with inflection points around .10 logits. In addition, the relationship between estimated classroom quality Rasch scores and associated standard errors for the Structural/Process specification were considerably more dispersed. As a result, in some instances individuals with the same estimated classroom Rasch score had markedly different standard errors, which indicated classrooms with identical levels of quality were measured with unequal precision. It was also useful to compare the coverage (i.e., range of the underlying latent continua) of the classroom quality Rasch scores across all models. This comparison was used to examine whether the multidimensional specifications of the measure improved the range of the underlying latent continuum covered by the items when compared to the two unidimensional specifications. The precision plot also indicated that neither of the multidimensional specifications of the measure added to the overall range of the underlying latent continuum covered by the items when compared to the two unidimensional specifications of the measure.

Structural Validity Evidence

To investigate the hypothesis that the Provisions for Learning/Teaching Interactions specification of the measure would best fit the data, the following analyses were undertaken: (1) the goodness-of-fit statistics were examined across all models, (2) convergent validity was examined using correlations between the Rasch scores for each dimension, and (3) discrepant case analyses were conducted to establish whether estimates for the multidimensional specifications of the ECERS-R led to substantive differences in the estimates for classroom quality.

Goodness-of-fit. The information-based fit indices (i.e., Akaike information criterion (AIC), Bayesian information criteria (BIC), sample adjusted Bayesian information criteria (aBIC), bias-corrected AIC (AICc), and the Bozdogan's Consistent Akaike information criterion (CAIC)) for all specifications of the measure are reported in Table 18. Model fit for the Provisions for Learning/Teaching and Interactions and Structural/Process specification can be directly compared to the 16-item unidimensional specification using all of the fit statistics provided in Table 18. However, because the 37-item unidimensional model contained additional items, only the BIC statistic can be used to compare this specification to all other specifications in the table (Raferty, 1986). Results showed that the 16-item unidimensional specification of the measure exhibited the best fit across all specifications of the measure.

Table 18. *Model Fit Statistics for All Model Specifications*

Model	AIC	AICc	BIC	aBIC	CAIC
Unidimensional 37-item	107047.80	107050.40	107268.70	107135.10	107310.70
Unidimensional 16-item	42258.81	42265.14	42600.6	42393.94	42665.60
Provisions for Learning/Teaching Interactions 2-Dimension	42360	42361	42486	42410	42510

Table 18. *Model Fit Statistics for All Model Specifications*

Model	AIC	AICc	BIC	aBIC	CAIC
Structural/Process 2-Dimension Within-Item Specification	44322	44323	44449	44372	44473

Note. The 37-item specification of the measure is not nested in the other models, so only the BIC statistic can be used to compare the fit of this model to the fit of other models. When comparing across the fit statistics, values that are smaller exhibit the best fit.

Comparison of Rasch Model Subscale Correlations. To examine the associations between dimensions both within and across model specifications, two sets of correlation coefficients were examined. The latent correlations were estimated for dimensions within each of the multidimensional specifications (i.e., latent correlations between the Provisions for Learning and Teaching and Interactions dimensions and latent correlations between the Structural and Process dimensions). These correlations were disattenuated from measurement error, and are reported in the upper diagonal of Table 19, which has been highlighted in gray. The other set of correlations reported in Table 19 are Pearson correlations, which were estimated from the estimated scores for classroom quality. It was necessary to view both sets of correlations because the latent correlations were only estimated within each multidimensional specification, and not between the different model specifications. Results showed that the Provisions for Learning/Teaching Interactions specification of the measure yielded a latent correlation of r = .66, between the Provisions for Learning and Teaching and Interactions dimensions and r = .55 for the Pearson correlation coefficients of the classroom quality estimates. The Structural/Process specification of the measure yielded a latent correlation of, r = .14, between the Structural and Process dimensions and r = .03 for the Pearson correlation coefficients of the classroom quality estimates. Looking across all model specifications, all of the Pearson correlation coefficients were positive and in the expected directions. Most coefficients were greater than .5, which indicated moderate to strong associations between dimensions. Two pairs of correlations were

notably low: (1) the association between the classroom quality estimates from the Teaching and Interactions Dimension and the Structural Dimension (r = .16), and (2) the association between Structural and Process Dimension showed an association of, r = .03.

Table 19. *Correlations Between Rasch Classroom Quality Estimates*

Model	1	2	3	4	5	6
Unidimensional 37-item	1.00					
Unidimensional 16-item	.93	1.00				
Provisions for Learning[a]	.85	.92	1.00	.66		
Teaching and Interactions[a]	.77	.82	.55	1.00		
Within Item Structural[b]	.60	.65	.87	.16	1.00	.14
Within Item Process[b]	.70	.76	.48	.92	.03	1.00

Note. All model specifications are for the collapsed 4-category rating scale. a = Rasch dimensional score from the Cassidy et al (2005[a]) 2-dimension specification of the measure. b = Rasch dimensional score from the Cassidy et al (2005[b]) 2-dimension within-item specification of the measure.

Discrepant Case Analyses. These analyses examined the instances of classrooms which differed notably in their dimensional estimated classroom quality when compared to both of the unidimensional specifications of the measure. To establish this, two sets of discrepant case analyses were conducted. In the first set of discrepant case analysis, each classroom quality estimate was standardized, and the differences between dimensional estimates and the scaled estimates for the unidimensional classroom quality estimates were examined. The goal was to establish those cases which differed by more than 1 standard deviation on any dimensional estimate for quality when compared to their unidimensional quality estimates (Allen & Wilson, 2006). The second set of analyses calculated the sums of squares discrepancy indicator, DI_p [12] (Allen & Wilson, 2006). This was a metric for establishing the cases that differed notably for the combined dimensional estimates for each multidimensional specification of the measure.

[12] $DI_p = \sum_{d=1}^{2}(\bar{\theta} - \theta_d)^2$, where $\bar{\theta}$ is the average case estimate on the unidimensional specification of the measure of interest, and θ_d = the case estimate for each dimension, d.

The percent of discrepant cases for each dimension are reported in Table 20. Criteria for establishing a discrepant case was a difference of one standard deviation between the classroom quality estimates for the multidimensional estimates for quality when compared to the unidimensional classroom quality estimates (Allen & Wilson, 2006). Across all multidimensional estimates, the percent of discrepant cases ranged from 8.02% to 23.45% when classroom quality estimates for the sub dimensions were compared to the 37-item unidimensional specification of the measure. Across both multidimensional specifications of the ECERS-R, the Structural/Process specification exhibited the largest percentage of discrepant cases. When the comparison group was the 16-item unidimensional specification of the measure, the percent of discrepant cases ranged from 2.67% to 18.52%. Again, the Structural/Process 2-dimension within-item specification of the measure exhibited the largest percentage of discrepant cases. The cross-plots for the different estimates of classroom quality for the sub-dimensions when compared to the 37-item unidimensional measure are shown in Figure 4. Discrepant cases are highlighted in green. The plots show that the discrepant cases were split across most of the range of classroom quality estimates. This result indicated that, across the entire range of the classroom quality estimates, a reliance on the unidimensional measure both over- and under-estimated levels of classroom quality for a sizeable percentage of the sample. Similar results were observed when the discrepant cases were based on the 16-item unidimensional version of the measure, and the cross plots for these cases are found in Figure 5.

Table 20. *Percent of Discrepant Cases for Each Model Specification*

Model	% Discrepant With 37-item Unidimensional Measure	% Discrepant With 16-item Unidimensional Measure
Provisions for Learning	8.02	2.67
Teaching and Interactions	13.52	10.14
Structural	23.45	18.52
Process	17.46	15.35

Table 20. *Percent of Discrepant Cases for Each Model Specification*

Model	% Discrepant With 37-item Unidimensional Measure	% Discrepant With 16-item Unidimensional Measure

Note. N = 1400, and is rounded to the nearest 100 per ECLS-B data reporting restrictions.

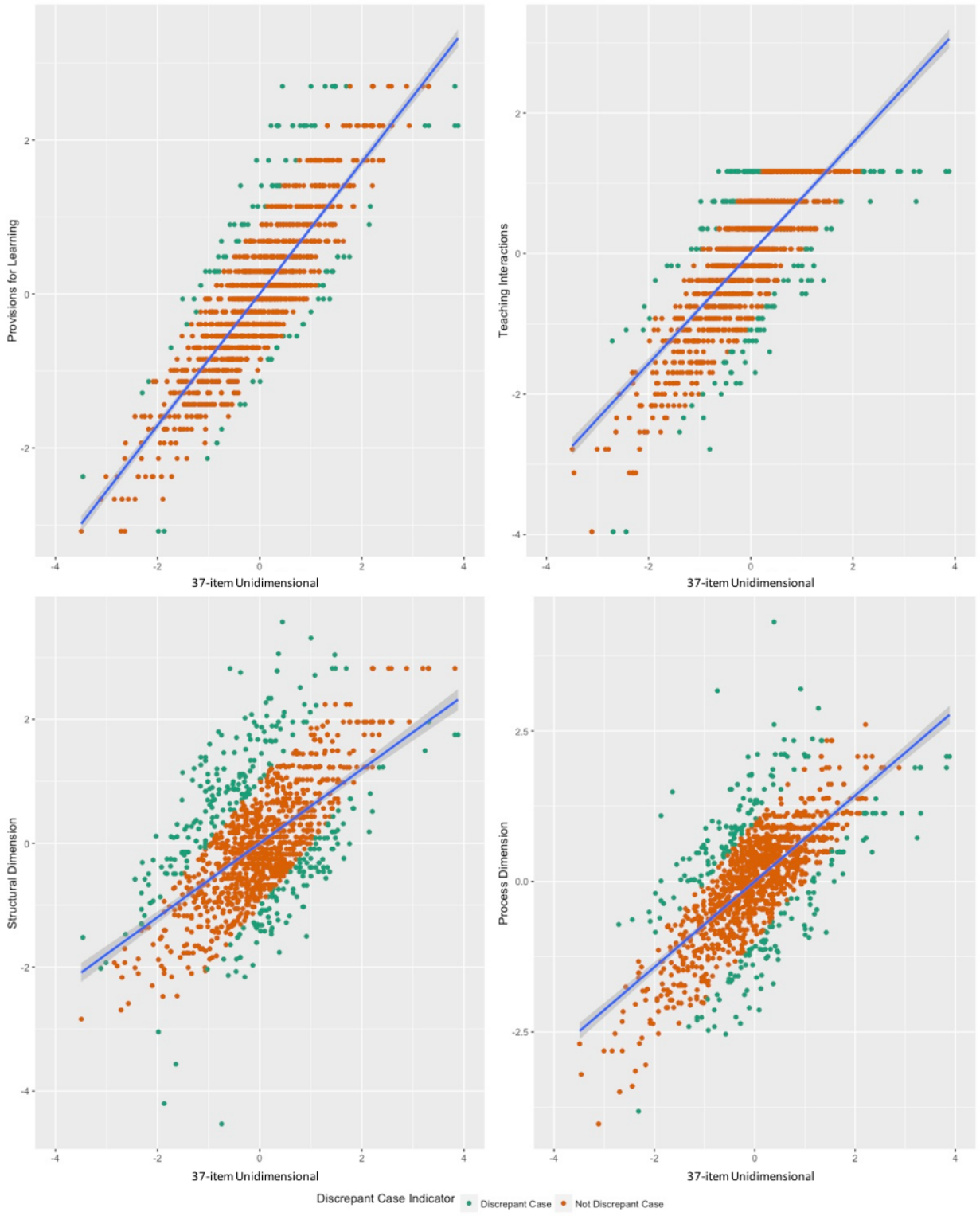

Figure 4. Discrepant Cases for each dimension of the two multidimensional specifications against scores on the 37-item unidimensional measure. Points in green indicate cases that differ by more than one standard deviation from their classroom quality on the dimension.

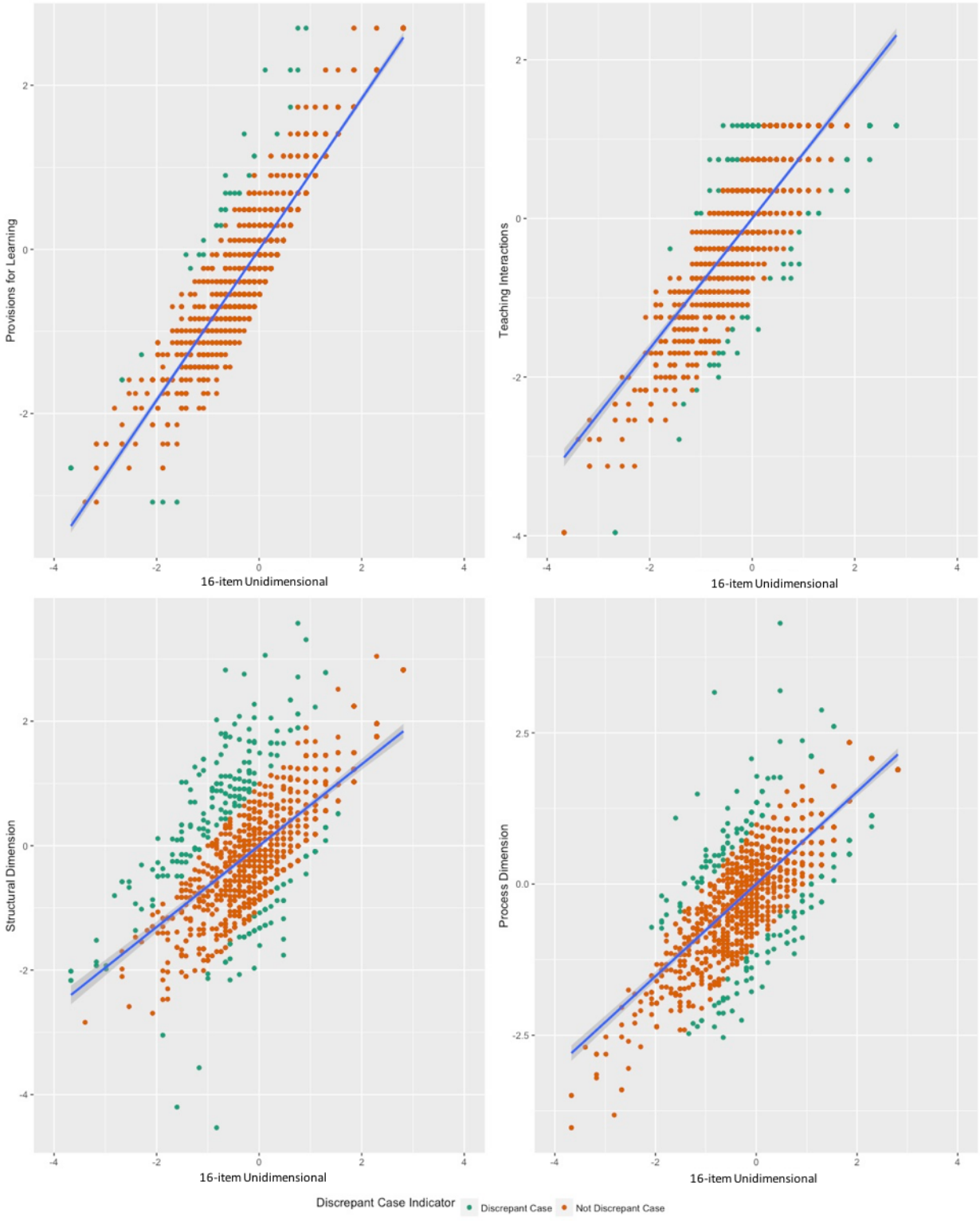

Figure 5. Discrepant Cases for each dimension of the two multidimensional specifications against scores on the 16-item unidimensional measure. Points in green indicate cases that differ by more than one standard deviation from their classroom quality on the dimension.

Next, the number of discrepant cases were compared across each multidimensional specification of the measure using the sums of squares discrepancy indicator, DI_p (Allen & Wilson, 2006). The goal of these analyses was to establish discrepant cases across both dimensions. This differs from the previous discrepant case analysis in that it accounts for classroom quality estimates for both dimensions simultaneously, instead of the one-to-one comparison. Results for these analyses showed that, when both dimensions for each specification were considered simultaneously, the percent of discrepant cases for the Provisions for Learning/Teaching and Interactions 2-dimension specification was $DI_p = 55.6\%$, and the percent of discrepant cases for the Structural/Process 2-dimension within-item dimensional speciation was $DI_p = 49.72\%$. The DI_p values were identical for both multidimensional specifications of the measure when the 16-item unidimensional specification of the measure was used as a reference.

Summary of Structural Validity Evidence. In sum, the structural validity evidence showed that both multidimensional specifications of the measures uniquely measured multiple domains, over and above the estimates of classroom quality that were produced for both unidimensional specifications of the measure. These results partially support the hypothesis that the Provisions for Learning/Teaching and Interactions specification of the measure was likely the best structural representation of the measure. However, when comparing the two multidimensional specifications of the measure, the Structural/Process specification yielded a higher percentage of discrepant cases, which indicated that this specification was more productive for measuring multiple dimensions of quality.

External

In order to provide evidence to support the hypotheses for the responsiveness of the measure, the following metrics from Wolfe & Smith (2007) were examined: (1) the Rasch

Person Strata Indices, (2) descriptive statistics for both the item difficulties and classroom quality estimates (i.e., Rasch scores), (3) the person item maps and the correlations between the classroom quality estimates, and (4) the total score on the Arnett Caregiver Interaction Scale.

The Rasch Person Strata Indices. The Rasch Strata Indices provided information to support the hypothesis that the measure would be able to distinguish between two to four statistically unique levels of classroom quality. Both the Rasch Person Separation (**G**) and Person Strata (**H**) indices are reported in Table 21 below. The results showed that, across all models, there was considerable variation in the number of performance levels captured by the dimensions in each measure. It is not surprising that the **G** and **H** indices for the Provisions for Learning, Teaching and Interactions, and Structural and Process dimensions are much smaller than both the 37 and 16-item unidimensional specifications for the measures, as a major factor in the calculation of these statistics is the number of items in the measure. The **G** and **H** statistics showed that the 37-item specification of the measure was able to distinguish between four to five performance levels, while the 16-item specification of the measure was able to distinguish between three to four performance levels, and both the Provisions for Learning/Teaching and Interactions, and Structural/Process dimensions were all able to distinguish between two to three performance levels.

Table 21. *Rasch Person Strata Indices*

Model	# of Items	Person Reliability	G	H
Unidimensional	37	.93	3.64	5.19
Unidimensional	16	.89	2.84	4.13
Provisions for Learning[a]	8	.84	2.29	3.39
Teaching and Interactions[a]	8	.77	1.83	2.77
Structural[b]	11	.81	2.06	3.09

Table 21. *Rasch Person Strata Indices*

Model	# of Items	Person Reliability	G	H
Process[b]	12	.83	2.21	3.28

Note. a = a dimension from the 2-dimension specification of the measure. b = items from the 2-dimension within-item specification of the measure. EAP = Expected A Posteriori (EAP) Person G = Person Separation Index, H = Person Strata Index.

Descriptive Statistics and Overlap for Item and Classroom Quality Estimates. The average item difficulty[13] for each specification of the measure varied slightly across model specifications. Comparing across all models, both unidimensional specifications of the measure contained items that were, on average, more difficult to endorse than the other model specifications, with the mean of the item difficulty calibrations for the 37-item unidimensional specification, $\overline{x}$ = -2.43, and $\overline{x}$ = -1.88 for the 16-item unidimensional specification. The mean of the item difficulty calibrations for the Provisions for Learning/Teaching and Interactions specification of the measure was $\overline{x}$ = -2.83, while the average of the item difficulties for the Structural/Process specification was $\overline{x}$ = -2.36. Further, the difference between the average classroom quality estimates ($\bar{\theta}$) and the item difficulty calibrations ($\overline{x}$) are reported in Table 22. The table demonstrates that, across all model specifications, there was a clear mismatch between the difficulty of the items and the overall classroom quality level in the sample. The difference in means across all model specifications was greater than 1 logit, which indicated the measure was likely too easy for this sample of classrooms.

Table 22. *Average Classroom Quality Rasch Estimates for Each Dimension*

Model	$\bar{\theta}$	SD	Min	Max	$\overline{x}$ - $\bar{\theta}$
Unidimensional 37-item	.29	.90	-2.37	2.86	-2.72
Unidimensional 16-item	1.37	1.29	-3.86	5.01	-3.25

[13] For the purposes of estimation, the classroom quality estimates were anchored as opposed to the item estimates. The choice of whether the "person estimates" or the item estimates are anchored is arbitrary.

Table 22. *Average Classroom Quality Rasch Estimates for Each Dimension*

Model	$\bar{\theta}$	SD	Min	Max	$\bar{x}$ - $\bar{\theta}$
Provisions for Learning	1.15	1.38	-3.09	4.87	-3.98
Teaching and Interactions	.02	1.81	-7.14	2.13	-2.85
Structural	.73	1.26	-4.96	5.23	-3.09
Process	1.26	1.45	-4.58	7.50	-3.62

Note. N = 1400, and is rounded to the nearest 100 per ECLS-B data reporting restrictions.

Finally, coverage statistics for the item difficulty estimates in relation to classroom quality estimates are reported in Table 23. These statistics are useful for understanding the percent of the sample that were covered by item difficulty calibrations. The coverage statistics were calculated using the minimum and maximum values of the threshold calibrations for the rating scale. Across all specifications of the measure the thresholds for the items sufficiently covered the classrooms.

Table 23. *Percent of Overlap Between Item Difficulty Estimates and Classroom Quality Rasch Scores*

Model	%Coverage
Unidimensional 37-item	> 95
Unidimensional 16-item	> 90
Provisions for Learning	> 95
Teaching Interactions	> 95
Structural	> 90
Process	> 90

Note. N = 1400, and is rounded to the nearest 100 per ECLS-B data reporting restrictions.

Examining the Wright Maps. The Wright Maps for each model specification are found in Figures 6-9. In addition, Wright Maps for the threshold calibrations for the easiest and most difficult items in each specification of the measure are provided for context in Appendix A. These maps plot the estimates for the average item difficulty calibrations relative to the classroom quality estimates. The estimates for both are on the same logit scale, which aids in interpretation on a number of fronts. These figures are especially useful in conjunction with

Table 23, which provides the coverage statistics for each item, as well as Tables 14, which provides the average item difficulty parameter estimates in order of difficulty to endorse.

The Wright Map for the 37-item unidimensional specification of the measure is found in Figure 6. The map showed that, for the majority of the classrooms in the sample, the items were too easy to endorse. Most of the estimates for the average item difficulty calibrations hovered around -1.5 to 3, with the largest concentration of average item difficulty estimates located at around -1.5 to -2 logits. Within this range of the construct, the items exhibited a high degree of construct saturation, meaning items differentiated minute differences in classroom quality at approximately the same point along the underlying continuum. However, ideally item calibrations should have been distributed across the entire range of the underlying latent continua. In the case of this specification of the measure, the items did not span the entire range of the latent continua. In general, items reflecting Personal Care were the most difficult to endorse, with "Meals/Snacks" (d = -1.16), "Diapering/Toileting" (d = -1.43), "Nap" (d = -1.51), "Safety Practice" (d = -1.66) and "Health Practice" (d = -1.75). Items pertaining to Interactions were the easiest to endorse, with "Staff-child Interactions" (d = -3.64), "Interactions among children" (d = -3.57), "Greeting/Departing" (d = -3.5), and "Encouraging Children to Communicate" (d = -3.39). Items pertaining to Activities, Space and Furnishings, Program Structure and Parents and Staff were interspersed with no discernable patterns in the mid-range of the measure. Finally, the majority of the estimated classroom quality scores were considerably higher than the average estimated average item difficulty calibrations, which indicated a ceiling effect in the measure.

The 16-item unidimensional specification of the measure is shown in Figure 7. The map shows a similar story as the 37-item specification in terms of the overall difficulty of the items in

relation to the estimated classroom quality scores. Again, the items in this specification of the measure were too easy for most of the sample. In examining the average item difficulty calibrations in the measures there were two distinct clusters of items that fell in the range of -2.5 to -2.25 and -.7 to -.5, respectively. Within these clusters, the average item difficulty calibrations were so closely clustered together that minute differences in classroom quality could be estimated. However, these clusters appeared as two distinct regions in the map, which indicated a gap in the measurement properties between the two clusters. Again, item calibrations did not adequately measure classrooms with high levels of classroom quality. The only discernable pattern in item difficulty were for items that were easy to endorse, as again items from the Interaction subscale were easiest to endorse, with "Staff-child Interactions" (d = -3.64), "Interactions among children" (d = -3.57), and "Greeting/Departing" (d = -3.5). The most difficult item in the measure was “Nature and Science,” which had a (d = -1.5). Items pertaining to Activities, Space and Furnishings, Program Structure and Parents and Staff were interspersed throughout the rest of the measure. The majority of the estimated classroom quality scores were considerably higher than the average estimated item difficulty calibrations, which indicated a ceiling effect in the measure.

The Wright Map for the Provisions for Learning/Teaching Interactions specification of the measure is displayed in Figure 8. Two distinct clusters of item calibrations can be found around -4 and -2. Items from the Provisions for Learning dimension were the more difficult to endorse than items from the Teaching and Interactions dimension. Items from this specification of the measure were dispersed more evenly across the range of about -4 to 0 logits, which still only covered classrooms with relatively low levels of classroom quality. Similarly, to the two unidimensional specifications, the items pertaining to classroom interactions were the easiest to

endorse, with "Informal use of language" (d = -2.57), "Discipline" (d = -2.58), "Group Time" (d = -2.65), "Interactions Among Children" (d = -3.2), and the "Staff-child Interactions" (d = -3.29) item. The average item difficulty calibrations for both the Provisions for Learning and Teaching and Interactions dimensions covered only a very minimal range of the estimated classroom quality measures. The majority of the estimated classroom quality scores were considerably higher than the average estimated item difficulty calibrations, which indicated a ceiling effect in the measure.

The Wright Map for the Structural/Process specification of the measure is shown in Figure 9. The average item difficulty calibrations were more dispersed along the latent continua when compared to the other specifications of the measure, and generally fell in the range of -4 to 0 logits. The dispersion of the average item difficulty calibrations along the Y-axis in the Wright Map indicated that this specification of the measure contained more items that contributed meaningfully (i.e., uniquely) to the measurement of the constructs. However, the range of the construct covered by the items was only a narrow range of the entire continuum for classroom quality. Again, the items with content which focused on the interactions between children and teachers/staff were the easiest to endorse, with "Informal use of language" (d = -4.32), "Discipline" (d = -4.34), "Group Time" (d = -4.43), "Interactions Among Children" (d = -5.14), and the "Staff-child Interactions" (d = -5.25) item, all of which aligned with the Process dimension. The most difficult item to endorse was the “Nature and Science” (d = -.82). The majority of the estimated classroom quality scores were considerably higher than the average estimated item difficulty calibrations, which indicated a ceiling effect in the measure.

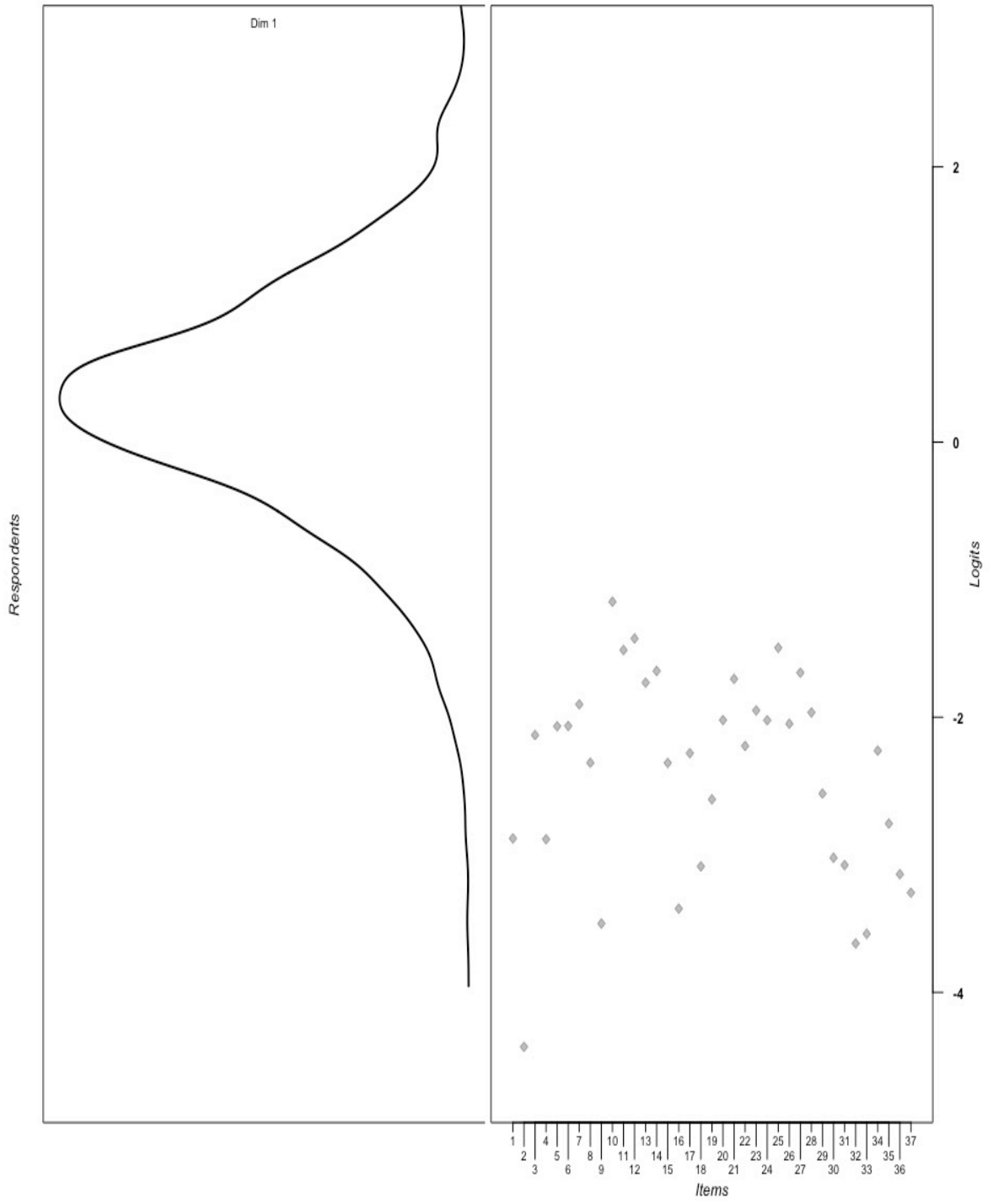

Figure 6. Person item map for the 37-item unidimensional specification of the measure. The number for each item reflects where that item fell within the measure, and can be triangulated with Table 2 above.

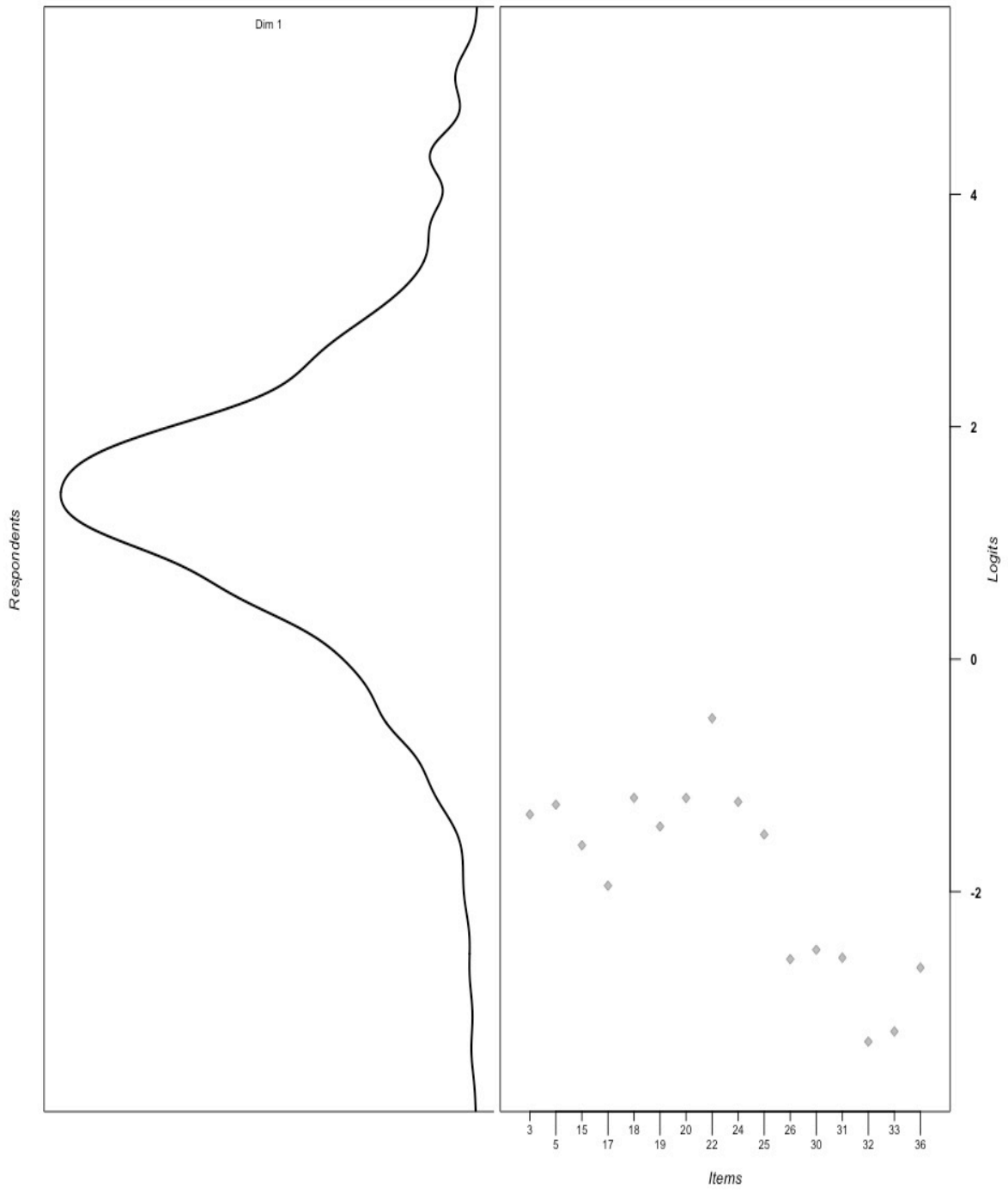

Figure 7. Person item map for the 16-item unidimensional specification of the measure. The number for each item reflects where that item fell within the measure, and can be triangulated with Table 2 above.

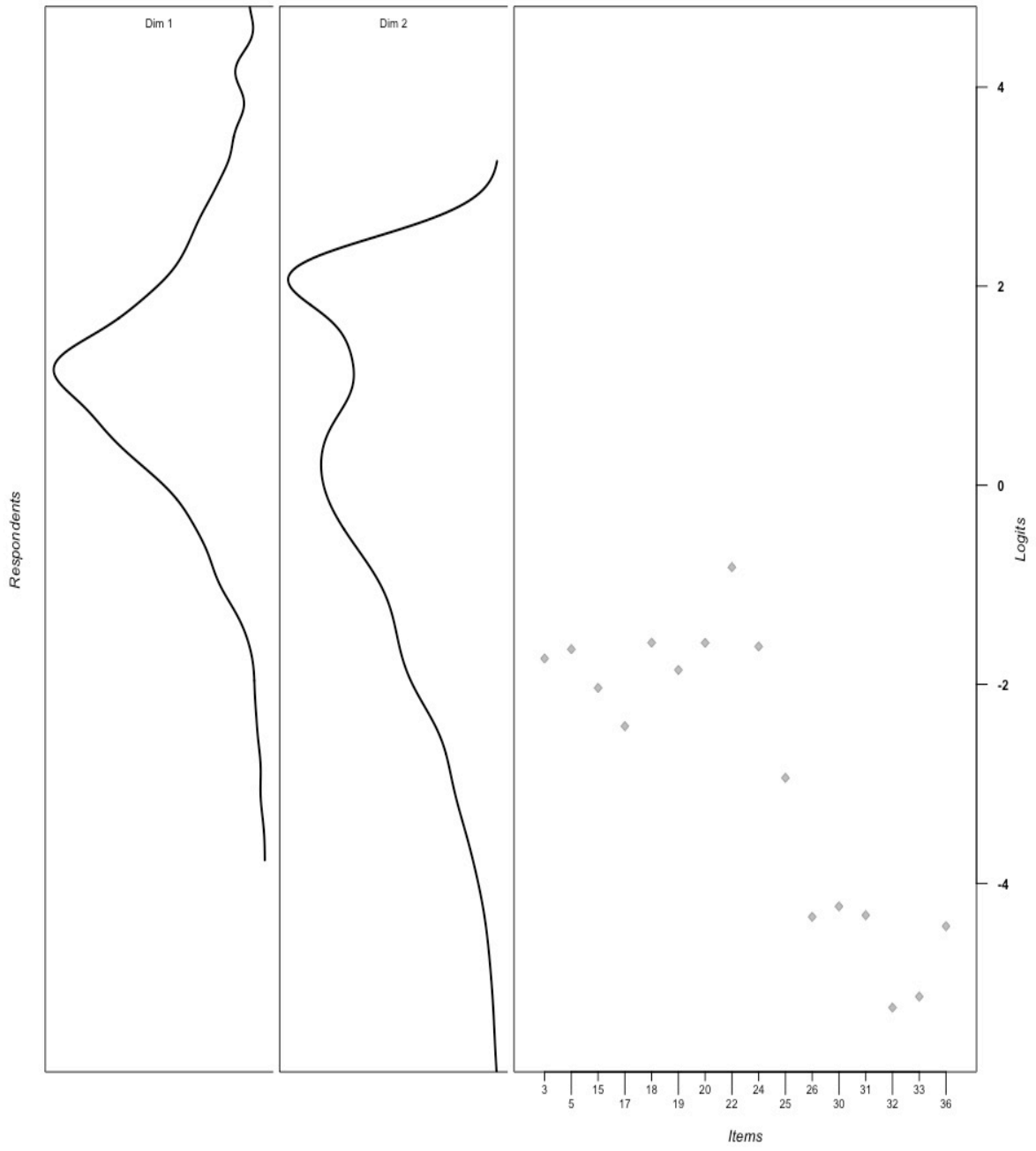

Figure 8. Person item map for the Provisions for Learning/Teaching Interactions two-dimensional specification of the measure. The number for each item reflects where that item fell within the measure, and can be triangulated with Table 2 above. Dim 1 = the Provisions for Learning dimension, while Dim 2 = the Teaching and Interactions dimension.

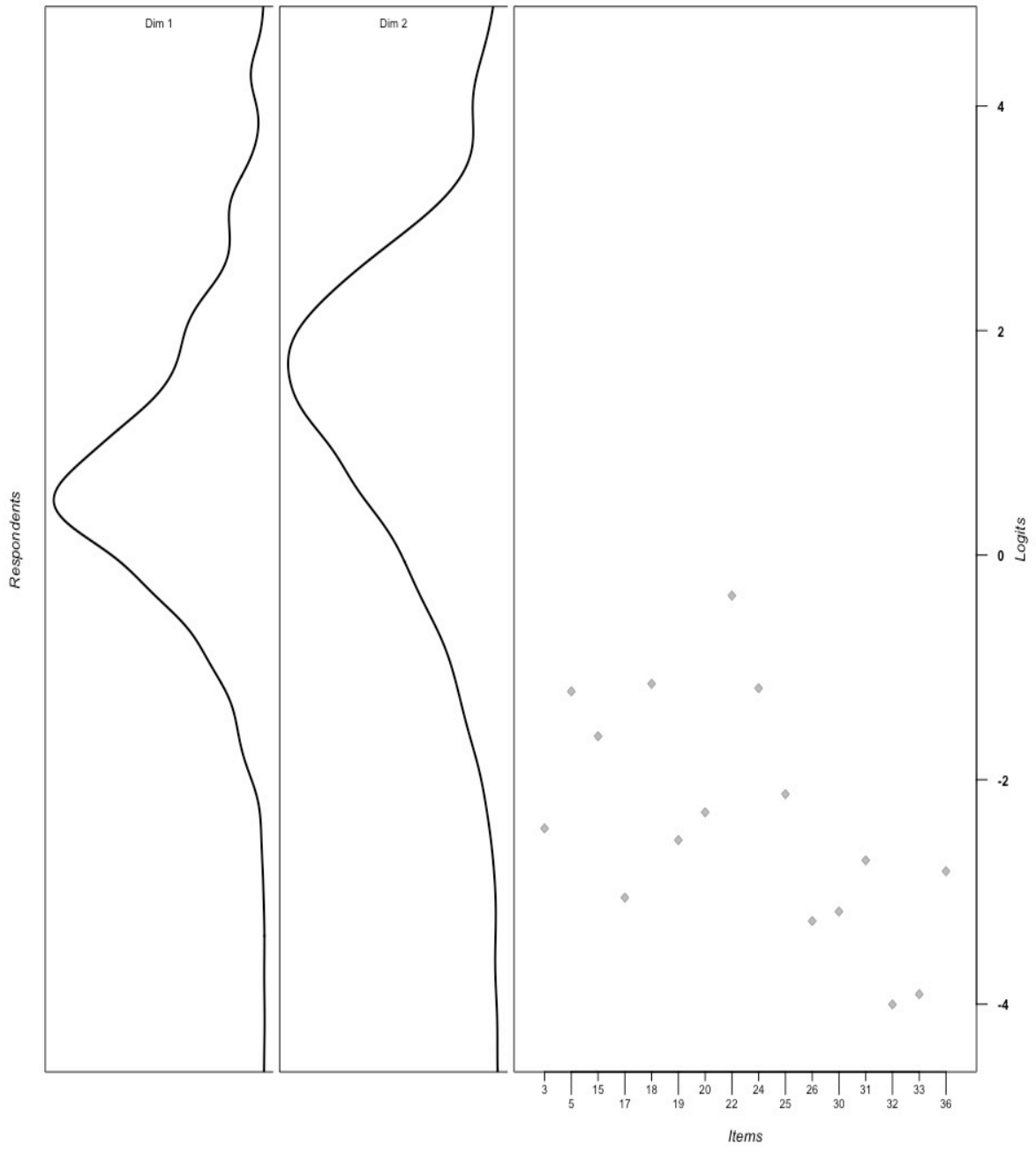

Figure 9. Person item map for the Structural/Process two dimensional within-item dimensional specification of the measure. The number for each item reflects where that item fell within the measure, and can be triangulated with Table 2 above. Dim 1 = the Structural dimension, while Dim 2 = the Process dimension.

Correlations with the Arnett Caregiver Interaction Scale. Pearson correlation coefficients for the associations between the estimates for classroom quality and the Arnett Caregiver Interaction Scale provided additional evidence for the concurrent validity of the measure. All of the associations with the Arnett Caregiver Interaction Scale were positive and in the expected direction. The strongest associations were with the Teaching and Interactions dimension, with $r = .74$, while associations for the Provisions for learning dimension were, $r = .45$. Both of the unidimensional specifications of the measure showed similar associations, with the 37-item unidimensional measure, $r = .62$, and for the 16-item unidimensional measure, $r = .66$. The Process dimension showed an association of, $r = .62$, with the Arnett Scale, while the Structural dimension was associated to a much lesser degree, with $r = .24$.

Summary of External Validity. These results largely supported the hypotheses pertaining to the external validity of the measure. First, the Rasch Person Strata indices showed that all specifications for the measure could distinguish between at least two statistically distinct levels of quality. However, the sensitivity of the measure differed depending on the number of items in the measure. For example, the measure was most responsive when all the available items were utilized in the 37-item unidimensional specification of the measure. Both multidimensional specifications of the measure were capable of detecting fewer statistically distinct levels of classroom quality. Further, a ceiling effect was identified in all specifications of the measure. Finally, correlations with the Arnett Caregiver Interactions scale were all positive and in the hypothesized direction. Further, these correlations were larger for the Teaching and Interactions and Process dimensions.

Predictive Validity

The final set of analyses were used to provide support to the hypothesis that no significant associations between the measure and child outcomes (i.e., children's reading and math skills) would be observed. Tables for all regression models can be found in Appendix B.

Associations with Children's Reading Outcomes. Regression analyses for the both the 37 item and 16 item unidimensional specifications of the measure showed no significant associations between estimates of classroom quality and children's preschool reading outcomes. Results for the multidimensional specifications of the measure showed that the Provisions for Learning dimension was positively associated with children's reading outcomes, with β = .03 (.00), t(38) = 15.26, p <.05. However, despite the statistical significance of the result, the R^2 value did not change with the addition of the classroom quality scores from $R^2 = .40$. In addition, both dimensions from the Structural/Process 2-dimension specification were associated with children's reading outcomes, with the Structural dimension β = .03 (.00), t(38) = 25.05, p <.05, and the Process dimension β = .02 (.00), t(38) = 39.38, p <.05. Again, the addition of the variables did not explain meaningful variance in the outcome, as the R^2 = .40 did not change with the addition of the classroom quality estimates.

Associations with Children's Math Outcomes. Regression analyses for the both the 37 item and 16 item unidimensional specifications of the measure showed no significant associations between estimates of classroom quality and children's preschool reading outcomes. Results for the multidimensional specifications of the measure showed that only the Provisions for Learning dimensions was associated with children's Math outcomes, β = .08 (.00), t(38) = 19.86, p <.05. However, the R^2 = .36 did not change from a prior model with all other statistical

controls. Consequently, the classroom quality estimates did not meaningfully account for the variance in children's math outcomes over and above the other statistical controls in the model.

Summary of Predictive Validity Evidence. Taken in conjunction, these results largely suggest that significant associations between classroom quality and child outcomes were only apparent for the multidimensional specifications of the measure. However, these associations did not account for a meaningful variance in student outcomes. Consequently, these results support the hypothesis for this study, that the measure was likely not associated with child outcomes.

Key Takeaways for the Psychometric Properties of the ECERS-R

In sum, results for each component of Messick's construct validity framework (1989) did not definitively suggest a particular specification of the ECERS-R was superior over other models that were examined. In particular, evidence pertaining to Messick's content component of validity showed that a higher proportion of items on the 37-item unidimensional specification of the measure were misfitting, while the multidimensional Structural/Process specification exhibited the lowest proportion of misfit, while the 16-item unidimensional and multidimensional Provisons for Learning/Teaching and Interactions exhibited the same proportion of misfitting items. These results suggested that the technical quality of the items within the Structural/Process specification were superior to the other models. Results pertaining to the substantive component of Messick's validity framework were consistent across all of the models examined, and showed that items with content generally centered on interactions between teachers and children were among the easiest to endorse in the measure. This was a result that suggested the theoretical conceptualizations underlying the expectations for how items on the ECERS-R performed did not conform to conceptualizations for how high-quality classrooms are described in the literature. Results that aligned with the generalizability validity

component showed that across all models examined, scores generated from the ECERS-R were more precise for classrooms with low to very low levels of classroom quality than for classrooms with average to above average levels of quality. In line with these results, analyses supporting the external component of validity showed that, across all models, the average item difficulty calibrations for items were easy and sufficiently covered the estimated scores for classroom quality for the classrooms with low levels of classroom quality. The Rasch person separation and strata indexes both indicated that undimensional versions of the measure were capable of detecting more levels of classroom quality in the sample than multidimensional specifications. However, in comparing the dispersion of the average item difficulty calibrations across all of the models there was evidence that the Structural/Process specification of the measure was superior, with average item difficulty calibrations that covered a wider range of the underlying continuum for classroom quality. Analyses for the supporting the structural component of Messick's framework indicated the presence of several discrepant cases in the sample. This was a result which indicated that ignoring the multidimensional specification of the measure might result in the estimation of classroom quality scores that either overestimate or underestimate the true levels of classroom quality in the sample. Finally, analyses investigating the predictive validity of the measure showed no significant associations between any model of the ECERS-R and child outcomes.

Discussion

Policymakers have increasingly come to view investments in expanding access to high-quality preschool as a critical way to optimize school readiness, bolster future student achievement, and strengthen the future workforce of the United States (Heckman, 2006; 2010). Federal and local governments are prioritizing this effort (Parker, Atchison & Workman, 2016; US Department of Education, 2016). As is the case with any large-scale policy effort that involves the expenditure of taxpayer dollars, policymakers have also implemented accountability efforts to ensure that preschool programs are effective (Schultz, 2015). The success of these accountability efforts is dependent on the ability to accurately measure classroom quality. As such, many states have co-opted extant observational measures of classroom quality into their accountability efforts – with the most utilized measure being the ECERS-R (Pianta, 2012). However, measures like the ECERS-R were not designed for the purposes of accountability. The process of collecting scores, making inferences about how the scores assess quality, and then making summative evaluative decisions involves several assumptions about the use of these measures for this application (Charalambous et al., 2012; Mashburn, 2017). These assumptions require an evidence base of rigorous and comprehensive psychometric information for this specific purpose (Charalambous et al., 2012; Gordon, 2013; 2015b; Hofer, 2012; Mashburn, 2017; Pianta, 2012; Sandilos & DiPerna, 2013). Yet, despite this, the use of measures like the ECERS-R in policy applications has not only persisted, but become increasingly high-stakes (Pianta, 2012). As such, it is now more important than ever that researchers focus on providing rigorous psychometric information to support or invalidate the use of this measure in various program and policy applications.

This dissertation provides the first comprehensive account of the psychometric properties of the ECERS-R, utilizing the ECLS-B dataset, a nationally representative sample of early childhood environments. As such, the goal of these analyses was to provide both researchers and policymakers with critical psychometric information that can be used to either support or disprove a range of claims about the validity of the ECERS-R for its varied uses in policy applications (Charalambous et al., 2012). To organize this evidence, this dissertation adopted Wolfe and Smith's (2007) Rasch validity framework, which ensured that analyses provided comprehensive psychometric information about each component of Messick's seminal (1989) construct validity framework. Further, several specifications of the ECERS-R have been both hypothesized and utilized in the literature. In some applications, the ECERS-R is used as a global measure of classroom quality (Gordon, 2013). Other applications of the ECERS-R consider it as a two-dimensional measure of classroom quality, with dimensions measuring Provisions for Learning and Teaching and Interactions (Cassidy et al., 2005a; Perlman et al., 2004; Sakai et al., 2004) or Structural and Process quality (Cassidy et al., 2005b). Each of these specifications of the ECERS-R necessitated their own comprehensive set of psychometric analyses. Findings pertaining to the structural validity of the measure suggested that for some classrooms, using the ECERS-R as a multidimensional measure led to substantively different estimates for quality when compared to using the measure as a global indicator for quality (i.e., unidimensional measure). However, these multidimensional specifications of the measure came with considerable tradeoffs with regard to the measurement properties of the instrument. In particular, multidimensional specifications of the measure exhibited less sensitivity than the unidimensional specifications of the measure, which indicated that these versions of the measure were capable of measuring fewer unique error strata (i.e., levels of classroom quality) in the sample. The analyses

also suggested that, regardless of the specification of the measure utilized (i.e., unidimensional or multidimensional), the ECERS-R appeared to be an extremely limited measure for classroom quality.

In this final chapter, results are discussed in three main sections. First, new insights into the dimensional structure of the measure are discussed. Next, a risk assessment frame is used to explain some of the limitations of the ECERS-R. This will be followed by a discussion of the implications of results for users and developers. The final section highlights the limitations of the analyses, as well as future directions for continued investigations into the psychometric properties of the ECERS-R.

New Dimensions in the Dimensional Debate for the ECERS-R

Most prior studies regarded as validity investigations for the ECERS-R have explored the factor structure of the measure. These studies have repeatedly shown that the ECERS-R measures two factors: a factor with items pertaining to Provisions for Learning, and another factor with items pertaining to Teaching and Interactions (Cassidy et al., 2005; Perlman et al., 2004; Sakai et al., 2004). Aside from the factor structure, information has lacked with regard to the psychometric properties for this specification of the measure, despite its repeated use in research studies. An aim of this dissertation was to thoroughly establish the psychometric properties for this specification of the measure. In addition, Cassidy and colleagues (2005b) have highlighted how items from the ECERS-R contain a blend of both process and structural indicators for quality, and in light of this, posited a Structural/Process dimensional specification for the measure that allows for some items to load onto both dimensions. However, to date, this specification of the measure had not been explored with quantitative methods. As such, an additional aim of this dissertation was to empirically establish whether the Structural/Process

specification of the measure was useful for measurement of classroom quality. Finally, a larger issue concerning the proper structural representation of the measure is whether different dimensional specifications of the measure could lead to qualitative or substantive differences in underlying levels of classroom quality, and whether these differences are large enough to warrant separate dimensional specifications of the measure. Results from these analyses shed light on several aspects concerning the aforementioned issues concerning the appropriate structural representation of the measure.

First, the discrepant case analyses revealed that both multidimensional specifications of the ECERS-R (i.e., Provisions for Learning/Teaching and Interactions and Structural/Process) were consistent with the hypothesis that scores for classroom quality estimated using the multidimensional specifications of the measure would capture substantive differences in classroom quality when compared to the use of unidimensional specifications of the measure. For example, there were classrooms in the sample with estimates for classroom quality on the Provisions for Learning/Teaching and Interactions and Structural/Process dimensions that differed by at least 1 standard deviation from the scores they obtained on both unidimensional specifications of the measure. For these classrooms, the dimensional estimates had the potential to reveal qualitatively or substantively different stories about classroom quality than if the measure was used as a global measure of classroom quality. However, results pertaining to which of the multidimensional specifications of the measure were preferable differed from what was expected. For example, the proportion of discrepant cases for the widely utilized Provisions for Learning/Teaching and Interactions specification of the measure were notably lower than the proportion of discrepant cases in the Structural/Process specification of the measure. This was an important finding. Given the history of the Provisions for Learning/Teaching and Interactions

specification of the measure in the literature, it was expected that the proportion of discrepant cases would have been notable enough to warrant the measurement of multiple dimensions for quality. Instead, the low proportion of discrepant cases in the sample suggested that this specification of the measure did not meaningfully differentiate multiple dimensions of quality to a degree which warranted the creation of two scores for classroom quality (i.e., scores for Provisions for Learning and Teaching and Interactions). These results also might shed light on the null or modest correlations observed between these dimensions and child outcomes (Auger et al., 2014; Burchinal et al., 2009; 2011; Gordon, 2013; Keys et al., 2013), while results from the present study were in line with the hypothesis of null associations between the ECERS-R and child outcomes, the findings bring new insights as to why these associations have not been observed in prior research. For example, the Provisions for Learning/Teaching and Interactions dimensions simply might not meaningfully distinguish between a high enough percentage of classrooms with truly average to high levels of classroom quality as to warrant that particular specification of the measure. Second, a notable pattern was observed in the Wright Maps of the average item difficulty calibrations for the Provisions for Learning/Teaching and Interactions specification of the measure. The item calibrations of both dimensions formed distinct clusters that were notably separated from one another along the continuum for classroom quality, with the Teaching and Interactions items measuring much lower levels of classroom quality than the items for the Provisions for Learning dimension. These results also showed that the Teaching and Interactions items were the easiest items in the measure. This was an unexpected result, given the research findings which have suggested that items with content pertaining to interactions between teachers and children should have been among the most difficult in the measure. For example, Perlman and colleagues (2004) found that when a group of teachers were

asked to select items from the ECERS-R that teachers perceived as items only "expert" teachers would score well on, they selected many of the Language and Interactions items. As such, it was expected that the average difficulty calibrations for these items would have differentiated between classrooms with higher levels of quality. Duncan (1984) has highlighted a major limitation in relying on factor analytic techniques to understand which items to retain in multidimensional measures, notably that the inter-item correlations and factor loadings are impacted by the average difficulty calibrations of the items. As such, if the underlying structure of the measure is truly unidimensional, these techniques can retain factors that are based on the difficulty of the items rather than true differences in constructs. Third, the EAP reliability statistics[14] from the generalizability validity analyses were higher for these dimensions when compared to the Structural/Process dimensions. The Wright Maps showed that items from the Provisions for Learning/Teaching and Interactions specification of the measure were less dispersed along the latent continuum than the items for the Structural/Process specification, which suggested that increased reliability for the Provisions for Learning/Teaching and Interactions specification came at the expense of coverage for the items along the latent continuum for classroom quality. This could have occurred through what Singh (2004) described as the bandwidth/fidelity problem, whereby researchers seek to maximize the correlations and reliability for items within a factor, which leads to a reduction in the range of the latent construct covered by the items. In sum, these results suggested that that a reliance on factor analytic techniques to establish this specification of the measure likely exaggerated differences between the dimensions.

[14] EAP means the Expected *a posteriori* reliability of the estimates for classroom quality, which is analogous to Cronbach's alpha.

Among the two multidimensional specifications of the ECERS-R that were examined, the Structural/Process specification, which allowed for within-item dimensionality, appeared to show the most promise. This result differed from the hypothesis in that it was expected that given the history and use of the Provisions for Learning/Teaching and Interactions specification of the measure, that the Structural/Process specification would have underperformed by comparison. Several key findings supported this conclusion. For example, this specification of the measure showed the largest proportion of discrepant cases. In addition, the correlations between these two dimensions, while in the expected direction with one another, were quite low. This suggested that this specification of the measure might more effectively differentiate between multiple dimensions of classroom quality. Another notable benefit to this speciation of the measure was that results from the generalizability analyses showed that this specification of the measure exhibited minimal item bias for any of the teacher or center characteristics that were examined. Further, the Wright Maps showed that the average difficulty calibrations of items for this specification were more evenly spaced across the range of the continuum captured by the items when compared to the Provisions for Learning/Teaching and Interactions specification, which has been used widely in the literature (Auger et al., 2014; Burchinal et al., 2009; 2011; Gordon, 2013; Keys et al., 2013; Perlman et al., 2004; Sakai et al., 2004). It has been more than a decade since Cassidy and colleagues (2005b) posited this specification of the measure, all the while researchers have continued to highlight how the issue of within-item dimensionality in the ECERS-R might be problematic for current conceptualizations about the dimensional structure of the measure (Gordon, 2013, Lambert, 2008). Yet, to date, these analyses represent the first attempt to operationalize the Cassidy et al (2005b) Process/Structural specification of the measure with psychometric methodology. The within-item dimensionality reflected in the items

of the ECERS-R should be considered in future studies that investigate the dimensional structure of the measure, as it may be the case that multidimensional specifications of the measure that require items to load onto only one dimension, like the widely utilized Provisions for Learning/Teaching and Interactions specification, run counter to how items within the measure actually function.

The ECERS-R as a Risk Assessment

The use of analysis techniques from the broader family of Rasch models represented a relatively innovative approach to understanding the functioning of the measure in the varied ways it used in both research and policy applications (Gordon, 2013; 2015b). A unique contribution of these analytic techniques included the generation of classroom quality scores that were linear, additive, interval-level, invariant, and hierarchical. Estimates for classroom quality and the difficulty of items were placed on the same scale to aid in direct comparison. Further, the scale reliability varied as a function of the level of the underlying construct(s) for classroom quality, which allowed for a more detailed assessment of the reliability of the measure at multiple levels of the construct. This modeling approach provided several insights into the overall difficulty of the measure for the classrooms in the sample, which have implications for the how the measure is used in policy applications.

Results from the Rasch analyses showed that, for all specifications of the ECERS-R, the measure was most able to differentiate between classrooms with relatively low levels of classroom quality. The Wright Maps, from the external validity analyses, succinctly displayed this issue. Through examining each map, it was clear that the average difficulty calibrations for the items were located at the lower end of the latent continuum for classroom quality. Within this lower range there was a high degree of construct saturation, which is to say items were clustered

very closely together with similar average difficulty calibrations. These results suggest what has long been suspected in the literature, which is that shorter versions of the ECERS-R, as implemented in the ECLS-B study at least, would suffice with little loss to the precision of the estimated levels of classroom quality or the range of the continuum covered by the average item difficulty calibrations (Cassidy et al., 2005a; Perlman, 2004; Scarr, 1994). Further, the distribution of classroom quality Rasch estimates was notably misaligned with the average difficulty calibrations for the items – this was the case regardless of the dimensional specification of the measure considered. The practical implication of this was that the measure could differentiate between classrooms with low levels of classroom quality with what would be analogous to microscopic precision. However, because the average item difficulty calibrations did not span the entire range of the underlying construct, classrooms with what the ECERS-R would describe as containing "Good" or "Excellent" levels of quality were measured poorly by the items. Evidence from the generalizability validity analyses triangulated this inference, as the Rasch reliability statistics and precision plots indicated that the measure provided the most reliable information for classrooms with low to average levels of classroom quality. The ECERS-R purports to be a global measure of quality, capable of distinguishing between "adequate" and "excellent" levels of quality (Harms, Clifford & Cryer, 1998). In fact, researchers have often utilized the measure to categorize classrooms into tiers of quality (Cassidy et al., 2005a; Perlman et al., 2004). Assumptions underlying these approaches are predicated on the idea that these tiers, or levels, of quality are substantively aligned with the semantic meaning of the words the test makers have used to identify categories of the rating scale (i.e., it ranges from 1 = adequate quality to 7 = excellent quality). In fact, researchers have often described their categorization of ECERS-R scores in line with how the rating scale is defined (e.g., Abner et al., 2013; Burchinal

et al., 2011; Le et al., 2015; Sakai et al., 2004). However, the results showed that while the invariance properties of the items that fit the Rasch model indicated that higher scores on items were associated with higher estimates for actual underlying levels of classroom quality (Rasch, 1960), this did not mean that these higher scores for quality were substantively associated with how the developers of the ECERS-R have worded the rating scale. As such, it was clear that descriptive terms like "excellent" or "adequate" quality must be considered with respect to the range of the quality that was captured by the items in the ECERS-R, as "excellent" classroom quality is relative to the most difficult items in the measure.

These results were expected for a number of reasons. First, similar patterns have been observed in other observational measures of quality. For example, Leventhal, Selner-O'Hagan, Brooks-Gunn, Bingenheimer and Earls (2004) demonstrated that the Home Observation Measurement of the Environment scale (HOME; Bradley & Whiteside-Mansell, 1998; Caldwell & Bradley, 1984) was most reliable for homes with low levels of quality. Second, Gordon and colleagues (2015b) had previously used a Partial Credit Model to show that the categories that are used to make up the rating scale for each item were disordered for every item in the ECERS-R, so it was not surprising that the issue generalized when other Rasch models were utilized to analyze this data. Further, this was also expected for the Provisions for Learning/Teaching and Interactions specification of the measure, which has been used widely in the literature. This solution was established through factor analysis in several articles (Cassidy et al., 2005a; Perlman et al., 2004; Sakai et al., 2005). The limitations of these analyses for measurement development are well known inside the measurement community, the most salient of which pertains to the bandwidth/fidelity tradeoff that occurs when these techniques are used for measurement development. Because factor analytic techniques seek to find solutions which

maximize the inter-item correlations, these methods tend to have the undesired effect of reducing the range of a particular construct captured by the items (Fan, 1998). However, it should be noted that while this issue was expected from a methodological standpoint, this does appear to be a general problem with the measure as a whole. This is because every specification of the ECERS-R that was examined contained item difficulty calibrations that were close together on the underlying continuum of latent classroom quality. As such, if any random subset of items was selected as a shorter form of the measure, similar bandwidth constraints for the items would have been observed.

Implications for Users

A known issue with the rating scale of the ECERS-R is that the categories of the rating scale are disordered for each item (Gordon, 2013; 2015b). Results from the present analyses replicated these results using an Andrich Rating Scale model, which assumed that the rating scale was the same across each item (e.g., that every item uses the same 7-point rating scale). An advantage of these models was that the results from analyses were more directly interpretable in applied contexts. This is because the Andrich Rating Scale Model assumes all items have the same number of steps, and the modeled distance between adjacent steps is consistent across items. Further, when the data fits the Rasch model, and the categories of the rating scale advance monotonically, total scores from the measure can be assured to be linear, additive, and interval-level, and implemented as interval data in parametric statistics (Andrich, 1978). However, violations of the ordering of the categories calls into question the simplicity with which scores from the ECERS-R have been used in the literature. Linacre (1999, pg. 111-112) describes the problem succinctly: “These average measures are an empirical indicator of the context in which the category is used. In general, observations in higher categories must be produced by higher

measures (or else we don't know what a 'higher' means). This means that the average measures by category, for each empirical set of observations, must advance monotonically up the rating scale, otherwise the meaning of the rating scale is uncertain. Consequently, any derived measures are of doubtful utility." As such, it is simply not the case that researchers can continue to use total scores from the measure, or its respective factors, without making attempts to recover the rating scale.

These analyses are the first attempt to recover the rating scale in a way that would allow for the use of total scores from the measure to be used in parametric statistics. Results showed that, by collapsing non-advancing categories into adjacent categories, a monotonic rating scale for the measure could be recovered (Linacre, 1999; 2002). As such, the 7-point rating scale became a 4-point rating scale. Researchers still continue to use these scores in the literature in contexts that have the potential to impact early childhood policy both within the United States (Burchinal et al., 2016; Coley et al., 2016; Dorman et al., 2017) and abroad (Brinkman, Hasan, Jung, Kinnell, Nakajima & Pradhan, 2016; Hasan, Brinkman, Jung, Kinnell, Nakajima & Pradhan, 2016; Mayer & Beckh, 2016; Sammons, Sylva, Hall, Siraj, Melhuish, Taggart & Mathers, 2017). However, item-level data for the ECLS-B remains available for researchers to utilize. In light of these results, as well as the results by Gordon and colleagues (2013; 2015b), researchers will need to scrutinize whether the use of total scores from the ECERS-R is helping or hindering advancements in the field of early childhood education. Those using this data for analyses will need to consider similar approaches to recovering the rating scale in order to use these measures appropriately.

Finally, both policymakers and researchers should consider the tradeoffs involved with using a multidimensional specification of the measure versus a unidimensional specification.

Analyses for the content component of validity showed that the item technical quality for items in the unidimensional specifications of the measure were mostly acceptable. While analyses for the substantive component of the validity of the ECERS-R did show that multidimensionality was important to consider, it comes with considerable tradeoffs. Both the Rasch Person Separation (**G**) and Person Strata (**H**) indices showed that the unidimensional specifications of the measure were able to differentiate one to two additional statistically distinct levels of classroom quality in the sample, meaning that the use of multidimensional specifications of the measure results in a considerable loss measure sensitivity. These results also have implications for policymakers, as in many state QRIS systems, tiers of quality are tied to scores on the ECERS-R. In many of these instances, the rating systems assume the ECERS-R is able to differentiate between more levels of quality than the results of these analyses showed (National Center on Child Care Quality Improvement, 2013; Tout, Chien, Rothenberg and Li, 2014)

Implications for Developers

The test makers of the ECERS-R are currently revising the measure (Clifford et al., 2012). To date, the revisions to the rating scale have only consisted of the elimination of the "stop-scoring" rule, which was previously recommended by Gordon and colleagues (2013). While the elimination of the stop-scoring rule has the desired effect of more fairly representing the levels of classroom quality in a classroom, there is little reason to believe elimination of this rule would rectify the issues of disordered categories in the rating scale. Results from the present study suggest a way in which the rating scale can be recovered if researchers have access to item level data for the measure. However, even in its recovered form, the rating scale for the ECERS-R exhibited issues that have ramifications for developers of the measure. In particular, researchers recommend that the estimates for each level of the rating scale not exceed 5 logits

between adjacent categories (Linacre, 1999). Across all specifications of the measure that were examined, the gap between the first and second category of the revised rating scale exceeded this criterion. Consequently, the first category of the rating scale represented a wider range of the underlying variable. Linacre (2002) describes that as the step calibrations between adjacent categories of the rating scale become wider apart, the information provided by items becomes more sensitive to the extreme scoring classrooms, and less sensitive to prototypical classrooms. These wide gaps in the between adjacent categories in the rating scale are suggestive of the presence of an unmeasured category that was not described sufficiently in the rating scale. A difficulty in how the rating scale is defined in the ECERS-R is that, despite the use of a consistent 7-point rating scale for each item, the rating scale itself is defined by the presence of key indicators in the classroom. These indicators are different across all of the items in the measure. As such, the results pertaining to disordered categories likely suggest a deeper structural issue in how the indicators for each item define each category in the rating scale, which is a result that has been confirmed by Gordon and colleagues (2015). Within a Rasch modeling approach, the relative ordering of the difficulty estimates for each indicator that defines each category of the rating scale should be ordered from least to most difficult. Unfortunately, given the number of indicators that define each category of the rating scale for each item, fulfilling this assumption is difficult to achieve without massive investments in new pilots to redevelop the indicators for the items. Results from the analyses suggested that the only real way to fix this problem would be to conceptualize the content and theoretical ordering of the indicators for each category of the rating scale for each item to represent less extreme differences in underlying quality. A simpler approach to defining the rating scale would be to utilize fewer categories, and to adopt an item prompt approach, similar to the CLASS, whereby

comprehensive descriptions for varying levels of classroom quality are described for each item in the measure. This would relieve some of the psychometric restrictions that need to be tested when using the current approach to defining the rating scale.

These analyses also shed light on another critical flaw of the measure. Namely, that the item hierarchy of the difficulty calibrations for the items were not intuitive or in line with what was expected. Developmental theory suggests that a hallmark of high quality classrooms is that they contain rich interactional episodes between teachers and children (Bronfrenbrenner & Morris, 1998; Vygotsky, 1979). Further, theoretical conceptualizations of classroom quality suggest that the purpose of things like activities or materials for learning in the classroom is to support and enhance these interactions between teachers and children (Cassidy et al., 2012; Cryer, 1999). However, substantive and external validity analyses showed that the Teaching and Interactions items of the ECERS-R were among the "easiest" on the measure. A reason for these results is likely that many of these items contain indicators that require observers to ask teachers about their interactions with children. As such, these items can be considered self-report measures, and are likely biased by teachers' desire to be viewed in a positive light. Developers of the ECERS-R would be wise to change these items to observational measures, similar to the CLASS. Doing so might result in these items becoming more "difficult" indicators of quality, and as such, expand the range of the latent construct covered by the items, which could help mitigate some of the issues with the limited bandwidth of the measure.

Further, many of the Activity items were among the most difficult to endorse. For example, the "Nature and Science", "Art", "Dramatic Play", and "Math" items were among the most difficult items across all specifications of the measure. It was hypothesized that these would be among the easiest, because the presence of activities and materials for learning in a classroom

is easily regulated. However, a closer inspection of the indicators for these items showed a clear progression from the presence of these materials in the classroom to teachers' use of these activities and materials to extend learning in other contexts. As such, within these items there was a hierarchy of behaviors that was theoretically predicated. For example, in most cases the first few categories of the rating scale were represented by indicators for structural quality, whereas later categories were represented by indicators for process quality. As such, within items that contained mixtures of both process and structural quality, the ordering of the difficulty of the indicators for each category of the rating scale fit with developmental conceptualizations, which suggest that process indicators of quality should distinguish classrooms with higher levels of quality. However, for developers these results might also suggest ways the measure could be revised to expand the range of classroom quality captured by the items. As an effective way to approach revisions to items would be to take the items from the ECERS-R that contain a mixture of both process and structural indicators, and to split these items according to the content of the indicators. For example, the Nature and Science item, which emerged as most difficult item for each 16-item specification of the measure could be split into a Nature and Science structural item, and a Nature and Science process item. Developers of the measure could then focus on developing new indicators for the rating scale to more fully define a more comprehensive range of the underlying continuum of quality for a Nature and Science process item. Redefining these items in this way could help with expanding the range of the latent continuum covered by the items.

However, it was still expected that the items from the Interactions and Language and Reasoning subscales of the ECERS-R would be among the most difficult on the measure. These results suggest that the item difficulty hierarchy of the measure is neither logical nor theoretically

predicated. This is problematic, given the varied uses of this measure in larger ECE accountability efforts (Pianta, 2012). In particular, for states using this measure for continuous improvement, the items from the measure should reflect a clear hierarchical set of behaviors that are both intuitive to teachers and logically indicative of increasing levels of quality. For example, based on these findings, does it make sense, from a practical standpoint, that the classrooms with the highest levels of quality receive high ratings for "Nature and Science", "Art", "Dramatic Play", and "Math" items, while classrooms with lowest quality struggle with these items, but do well on the interaction items? These are questions both policymakers and researchers should consider when using this measure to understand issues pertaining to preschool classroom quality.

Limitations and Future Directions

These analyses come with limitations. A strength of using the ECLS-B dataset for analyses was that the data was based on a nationally representative sample of children's early childhood experiences. However, as children were followed into early childcare, only those providers who agreed to participate in the study provided data. Consequently, analyses might be biased by selection effects. Another weakness of this study was specific to the analyses of child outcomes. Designers of the ELCS-B survey chose to create their own measures for children's cognitive skills. These measures were composites which consisted of items from a variety of available measures. While these composite measures have been examined with extensive psychometric analyses (Najarian, Snow, Lennon, Kinsey & Mulligan, 2007), there is the potential that associations with child outcomes would have looked different with more domain specific measures for child outcomes. Importantly, the Structural/Process specification of the measure was associated with child outcomes, albeit to a magnitude that was not meaningful in this sample, it might be the case that these new dimensions from the measure would be more

meaningfully associated with other domain specific assessments for child outcomes. Given that this study was the first to examine the associations between these dimensions and child outcomes using Rasch scores for classroom quality, there is no literature to draw on to predict how these associations might look in future samples. Another weakness in the analyses of child outcomes was that these data were situated inside of a multilevel ecological context; however, the sampling frame of ECLS-B did not allow for the use of multilevel modeling techniques. Finally, it should be noted that the Structural/Process specification of the measure used 16 of the 37 items that were available in the ECLS-B data set. The reason for this was to ensure that the items used for analyses were commensurate with the items that were used in the Cassidy et al. (2005a) Provisions for Learning/Teaching and Interactions specification. However, future analyses could explore how the inclusion of additional items from the ECERS-R might impact the measurement properties of the ECERS-R when the Structural/Process specification is utilized.

The current analyses exemplified the practical utility of adopting methods from item response theory, as the methods provided comprehensive information about the psychometric properties of the ECERS-R. To date, the use of item response models to explore the psychometric properties of early childhood classroom observational measures is still in its nascent stages (Gordon 2013; 2015). The Andrich Rating Scale model and the multidimensional random coefficients multinomial logit (MRCML) model are two of many psychometric models that could be used to analyze item-level data for the ECERS-R. The rationale for the selection of these models was to align the methodology of the study with how the measure was conceptualized and used in practice. For example, both of these models assumed that the rating scale is the same across all of the items. This, of course, is how the developers of the ECERS-R have conceptualized the rating scale (Harms et al., 1998). However, the use of different item

response models might help researchers glean insights into other psychometric properties of the measure that were not currently explored. For example, Gordon and colleagues (2013) utilized a Rasch Partial Credit Model (Masters, 1982) to show how severe the problem of disordered categories in the rating scale was for each item in the measure.

Researchers are increasingly expressing concern with regard to possible rater effects (i.e., rater severity) for early childhood observational measures (Kane, Kerr & Pianta, 2014; Pianta, 2012). This is because an assumption underlying the use of scores from these measures, in both policy and research applications, is that items should be rated the same irrespective of the observer. The concern with observational measures like the ECERS-R is that different observers might rate the same item with varying degrees of severity or leniency. As such, future analyses should utilize psychometric methods that allow researchers to understand whether estimates of classroom quality are biased due to rater effects. And there is good reason to believe the ECERS-R might display issues with rater severity, as researchers have highlighted how rater effects tend to be more apparent in measures that ask observers to make global judgements (Pianta, 2012). In the future, researchers might consider using psychometric models like a Facets model (Linacre & Wright, 2002) to investigate potential rater effects. In a Facets model the ordinal observations are conceptualized as the outcome of interactions between elements of the classroom Rasch measures, item difficulty, and rater severity or leniency. Further, these models also allow researchers to explore whether rater severity or leniency is biased towards certain teacher characteristics or program types (Engelhard, 2007). Results from these models would have important implications for policy and practice. For example, the discovery of widespread issues of rater bias would highlight the need for ongoing statistical monitoring of raters, and results from these analyses could be used by states for feedback or remediation efforts.

Finally, Harms and colleagues (1998) should be commended for permitting the reporting of item-level data in large-scale datasets like the ELCS-B. Revolutions in science are rare, and progress is often incremental (Kuhn & Hawkins, 1963). Incremental progress depends on openness, a willingness to share data (Borgman, 2012). However, it is often the case that the incentive structure in science favors novelty, and lackluster results can persist when the scientific community does not embrace an open source culture (Nosek, Spies & Motyl, 2012; Nosek et al., 2015). The obvious shifts in the policy landscape for early childhood programs towards an ethos of accountability is rightfully causing researchers to question the validity of observational measures for quality. It has been the availability of item-level data for the ECERS-R that has allowed researchers to conduct the kinds of analyses necessary to understand the unintended consequences for using this measure in policy applications (Gordon, 2013; 2015b). Consequently, researchers are increasingly sounding alarms about the validity of the ECERS-R (Goldstein & Flake, 2016; Votruba-Drzal & Miller, 2016), specifically. However, there is reason to believe that the issues documented with the ECERS-R would be evident in other measures for classroom quality. A concerning trend in the literature has been for researchers to postulate that the CLASS is a better observational measure of quality for early childhood classroom environments. To date, the evidence for this assumption has been weak. The validity investigations that exist consist almost exclusively of factor analyses (Hafen, Hamre, Allen, Bell, Gitomer & Pianta, 2015; Paro, Pianta, & Stuhlman, 2004; Malmberg, Hagger, Burn, Mutton, & Colls, 2010; Pakarinen, et al., 2010; Virtanen, Pakarinen, Lerkkanen, Poikkeus, Siekkinen & Nurmi, 2017). Further, associations between the CLASS and child outcomes have been small enough to raise real concerns about the practical implications of such small correlations if this measure is used in policy applications (Burchinal et al., 2009; 2012; Keys et al., 2013; Mashburn

et al., 2008). Yet, despite this, the CLASS is emerging as the formidable successor to the ECERS-R, and is increasingly being used for high-stakes summative policy decisions (Mashburn, 2017). Of course, there is a strong theoretical rationale that supports the idea that a more domain specific measure of classroom quality, which focuses exclusively on the interactions between children and teachers, would be superior to the ECERS-R (McCabe & Ackerman, 2007; Mashburn et al., 2008). However, only rigorous psychometric analyses of the item-level data from this measure could provide the kind of evidence necessary to demonstrate its superiority.

Summary

Limitations notwithstanding, this dissertation adds to the existing literature pertaining to the validity of the ECERS-R. This was the first study to utilize Rasch modeling techniques to provide information about the psychometric properties of the measure for each component of Messick's (1989) Construct Validity Framework. Results contribute to the literature by showing that the ECERS-R functioned best as an instrument capable of measuring relatively low-levels of classroom quality. Several specifications of the measure were examined, and results showed that multidimensional specifications of the ECERS-R led to substantively different scores for classroom quality than when the measure was used as a global measure of classroom quality. Further, a new multidimensional specification of the measure, which allowed for items to load onto both a Structural and Process dimension, was the most responsive multidimensional specification of the measure. However, the use of the ECERS-R as a multidimensional measure came at considerable cost to the sensitivity of the measure. Much more work remains, and future research should utilize additional models from item-response theory to illuminate additional psychometric properties of the measure.

References

Abbott-Shim, M., & Sibley, A. (1998). Assessment Profile for Early Childhood Programs: Research Edition II. Atlanta, GA: Quality Counts, Inc.

Abbott-Shim, M., Sibley, A., & Neel, J. Assessment profile for early childhood programs—research version, 1992. *Quality Assist, Atlanta, GA*.

Abner, K. S., Gordon, R. A., Kaestner, R., & Korenman, S. (2013). Does Child-Care Quality Mediate Associations Between Type of Care and Development?. *Journal of Marriage and Family*, *75*(5), 1203-1217.

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*(2-3), 162-172.

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, *72*(4), 547-573.

Aikens, N., Bush, C., Gleason, P., Malone, L., & Tarullo, L. (2016). *Tracking Quality in Head Start Classrooms: FACES 2006 to FACES 2014*. Mathematica Policy Research.

Allen, D. D., & Wilson, M. (2006). Introducing multidimensional item response modeling in health behavior and health education research. *Health Education Research*, *21*, 73-i84.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, *9*(1), 95-104.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike* (pp. 199-213). Springer New York.

American Educational Research Association, American Psychological Association, National

Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Andrich, D., De Jong, J. H. A. L., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. *Applications of Latent Trait and Latent Class Models in the Social Sciences*, 59-70.

Arnett, J. (1989). Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology, 10*, 541–552.

Barack Obama, President of the United States. (2013). State of the Union Address.

Barnett, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *The Future of Children*, 25-50.

Barnett, W. S. others, "The State of Preschool 2015: State Preschool Yearbook" (New Brunswick, NJ: National Institute for Early Education Research, 2016).

Belfield, C., Nores, M., Barnett, W. S., & Schweinhart, L. (2005). Updating the Benefit-cost Analysis of the High/Scope Perry Preschool Program through Age 40. *Educational Evaluation and Policy Analysis*, *27*(3), 245-262.

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.

Beller, E. K., Stahnke, M., Butz, P., Stahl, W., & Wessels, H. (1996). Two measures of the

quality of group care for infants and toddlers. *European Journal of Psychology of Education*, *11*(2), 151-167.

Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist*, *57*(2), 111.

Bodrova, E. L. E. N. A., Leong, D. J., Dickinson, D. K., & Neuman, S. B. (2006). Vygotskian perspectives on teaching and learning early literacy. Handbook of early literacy research, 2, 243-256.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078.

Bradley, R. H., & Whiteside-Mansell, L. (1998). Home environment and children's development: Age and demographic differences. In M. Lewis & C. Feiring (Eds.), *Families, Risk, and Competence* (pp. 133-157). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bredekamp, S. (1997). NAEYC issues revised position statement on developmentally appropriate practice in early childhood programs. *Young Children, 52*(2), 34-40.

Bredekamp, S., & Copple, C. (1997). Developmentally appropriate practice in early childhood education. *Washington, DC: National Association for the Education of Young Children*.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*(1), 87-100.

Brinkman, S. A., Hasan, A., Jung, H., Kinnell, A., Nakajima, N., & Pradhan, M. P. (2016). The role of preschool quality in promoting child development: evidence from rural Indonesia. *World Bank Policy Research Working Paper*, 7529.

Bronfenbrenner, U., & Morris, P. A. (1998). The ecology of developmental processes.

Brownell, M. D., Nickel, N. C., Chateau, D., Martens, P. J., Taylor, C., Crockett, L., Katz, A. Sarkar, J., Burland, E. & Goh, C. Y. (2015). Long-term benefits of full-day kindergarten: a longitudinal population-based study. *Early Child Development and Care*, *185*(2), 291-316.

Buell, M., Han, M., & Vukelich, C. (2016). Factors affecting variance in Classroom Assessment Scoring System scores: season, context, and classroom composition. *Early Child Development and Care*, 1-14.

Buettner, C. K., Andrews, D. W., & Glassman, M. (2009). Development of a student engagement approach to alcohol prevention: The Pragmatics Project. *Journal of American College Health*, *58*(1), 33-38.

Burchinal, M. R., Cryer, D., Clifford, R. M., & Howes, C. (2002). Caregiver training and classroom quality in child care centers. *Applied Developmental Science, 6*(1), 2-11.

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. *Quality Measurement in Early Childhood Settings*, *11*.

Burchinal, M. R., Kainz, K., Cai, K., Tout, K., Zaslow, M., Martinez-Beck, I., & Rathgeb, C. (2009). Early care and education quality and child outcomes. *Washington, DC: Child Trends*.

Burchinal, M. R., Magnuson, K., Powell, D., & Hong, S. S. (2015). Early childcare and education. In M. H. Bornstein & T. Leventhal (Eds.). Ecological settings and processes in developmental systems (Vol. 4). In Handbook of child psychology and developmental science, 7th edition (Editor-in-Chief., R. M. Lerner) (pp. 223-267). Hoboken, NJ: Wiley

Burchinal, M. R., Zaslow, M., & Tarullo, L. (2016). *Quality Thresholds, Features, and Dosage in Early Care and Education: Secondary Data Analyses of Child Outcomes*. Wiley-Blackwell.

Burchinal, M. R., Xue, Y., Auger, A., Tien, H. C., Mashburn, A., Peisner-Feinberg, E., Cavadel, E.W., Zaslow, M. & Tarullo, L. (2016). III. TESTING FOR QUALITY THRESHOLDS AND FEATURES IN EARLY CARE AND EDUCATION. Monographs of the Society for Research in Child Development, 81(2), 46-63.

Burkam, D. T. (2013). Educational Inequality and Children: The Preschool and Early School Years. *The Economics of Inequality, Poverty, and Discrimination in the 21st Century [2 Volumes]*, 381.

Caldwell, B., & Bradley, R. (1984). *Home Observation Measurement of the Environment*. Little Rock: University of Arkansas.

Campbell, F. A., Pungello, E. P., Miller-Johnson, S., Burchinal, M., & Ramey, C. T. (2001). The development of cognitive and academic abilities: growth curves from an early childhood educational experiment. *Developmental Psychology*, *37*(2), 231.

Carneiro, P. M., & Heckman, J. J. (2003). Human capital policy.

Cassidy, D. J., Hestenes, L. L., Hansen, J. K., Hegde, A., Shim, J., & Hestenes, S. (2005). Revisiting the two faces of child care quality: Structure and process. *Early Education and Development*, *16*(4), 505-520.

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*(3), 345-360.

Child Trends (2014). Family structure: indicators on children and youth. *Child Trends Data Bank*.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Civic Impulse. (2017). H.R. 1532 — 112th Congress: Race to the Top Act of 2011. Retrieved from https://www.govtrack.us/congress/bills/112/hr1532.

Clarke-Stewart, K. A., Vandell, D. L., Burchinal, M., O'Brien, M., & McCartney, K. (2002). Do regulable features of child-care homes affect children's development?. *Early Childhood Research Quarterly*, *17*(1), 52-86.

Clifford, R. M., Reszka, S. S., & Rossbach, H. G. (2010). Reliability and validity of the early childhood environment rating scale. *Retrieved September*, *30*, 2013.

Clifford R, Sideris J, Neitzel J, Abuchaim B. A new scoring approach for ECERS–R [PowerPoint slides] 2012 Retrieved from http://ers.fpg.unc.edu/presentations.

Coley, R. L., Votruba-Drzal, E., Collins, M., & Cook, K. D. (2016). Comparing public, private, and informal preschool programs in a national sample of low-income children. *Early Childhood Research Quarterly, 36,* 91-105.

Colwell, N., Gordon, R. A., Fujimoto, K., Kaestner, R., & Korenman, S. (2013). New evidence on the validity of the Arnett Caregiver Interaction Scale: Results from the early childhood longitudinal study-birth cohort. *Early Childhood Research Quarterly*, *28*(2), 218-233.

Congdon, P. J., & McQueen, J. (2000). Unmodeled rater discrimination error. *Objective Measurement: Theory into practice,* (5), 165-180.

Congress, U. S. (1994). Goals 2000: Educate America Act. *Public Law*, 103-227.

Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A

motivational analysis of self-system processes.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology, 78*(1), 98.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin,* 52(4), 281.

Crosnoe, R., Leventhal, T., Wirth, R. J., Pierce, K. M., & Pianta, R. C. (2010). Family socioeconomic status and consistent environmental stimulation in early childhood. *Child Development*, *81*(3), 972-987.

Crosnoe, R., Purtell, K. M., Davis-Kean, P., Ansari, A., & Benner, A. D. (2016). The selection of children from low-income families into preschool. *Developmental Psychology*, *52*(4), 599.

Cryer, D., and Burchinal, M. R. (1997). Parents as child care consumers. *Early Childhood Research Quarterly, 12*, 35–58.

Curby, T. W., Downer, J. T., & Booren, L. M. (2014). Behavioral exchanges between teachers and children over the course of a typical preschool day: Testing bidirectional associations. *Early Childhood Research Quarterly*, *29*(2), 193-204.

Curby, T. W., Rimm-Kaufman, S. E., & Ponitz, C. C. (2009). Teacher–child interactions and children's achievement trajectories across kindergarten and first grade. *Journal of Educational Psychology*, *101*(4), 912.

Davis, E. A., & Miyake, N. (2004). Explorations of scaffolding in complex classroom systems. *The Journal of the Learning Sciences*, *13*(3), 265-272.

de Kruif, R. E. L., McWilliam, R. A., Ridley, S. M., & Wakely, M. B. (2000). Classification of

teachers' inter- action behaviors in early childhood classrooms. *Early Childhood Research Quarterly, 15*, 247 – 268.

DeGangi, G. A. (1995). Infant/Toddler Symptom Checklist: A screening tool for parents. Psychological Corp.

Dorman, R. L., Anthony, E., Osborne-Fears, B., & Fischer, R. L. (2017). Investing in high quality preschool: lessons from an urban setting. *Early Years, 37*(1), 91-107.

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical.* Russell Sage Foundation.

Duncan, S. E., & De Avila, E. A. (1998). PreLAS 2000. *Monterey, CA: CTB/McGraw-Hill*.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P. & Sexton, H. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428.

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *The Journal of Economic Perspectives, 27*(2), 109-132.

Dunn, L. (1993). Proximal and distal features of day care quality and children's development. *Early Childhood Research Quarterly*, *8*(2), 167-192.

Dunn, L. M., & Dunn, L. M. (1997). PPVT-III: Peabody picture vocabulary test. Circle Pines, MN: American Guidance Service.

Early, D. M., Bryant, D. M., Pianta, R. C., Clifford, R. M., Burchinal, M. R., Ritchie, S., Howes, C. & Barbarin, O. (2006). Are teachers' education, major, and credentials related to classroom quality and children's academic gains in pre-kindergarten?. *Early Childhood Research Quarterly*, *21*(2), 174-195.

Emlen, A. (1999). From a Parent's Point of View: Measuring the Quality of Child Care, Portland

State University, Portland OR.

Engelhard, G. (2007). Differential rater functioning. *Rasch Measurement Transactions, 21*(3), 1124.

Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357-381.

Flanagan, K. D., & West, J. (2005). *Children born in 2001: First results from the base year of the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B)* (No. 4825). Mathematica Policy Research.

Goelman, H., & Guo, H. (1998). What we know and what we don't know about burnout among early childhood care providers. *Child and Youth Care Forum, 27*, 175-199.

Goldstein, H., & Healy, M. J. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 175-177.

Gordon, R. A. (2015a). Measuring Constructs in Family Science: How Can Item Response Theory Improve Precision and Validity?. *Journal of Marriage and Family, 77*(1), 147-176.

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.

Gordon, R. A., Hofer, K. G., Fujimoto, K. A., Risk, N., Kaestner, R., & Korenman, S. (2015b).

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*, *26*(8), 1086-1110.

Gorey, K. M. (2001). Early childhood education: A meta-analytic affirmation of the short-and long-term benefits of educational opportunity. *School Psychology Quarterly*, *16*(1), 9.

Gormley Jr, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology*, *41*(6), 872.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977, Winter). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*(4), 827-833.

Grolnick, W. S., & Ryan, R. M. (1989). Parent styles associated with children's self-regulation and competence in school. *Journal of Educational Psychology*, *81*(2), 143.

Guo, Y., Piasta, S. B., Justice, L. M., & Kaderavek, J. N. (2010). Relations among preschool teachers' self-efficacy, classroom quality, and children's language and literacy gains. *Teaching and Teacher Education, 26*(4), 1094-1103.

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms: Revisiting the factor structure and practical application of the classroom assessment scoring system–secondary. *The Journal of Early Adolescence, 35*(5-6), 651-680.

Hahs-Vaughn, D. L. (2005). A primer for using and understanding weights with national datasets. *The Journal of Experimental Education*, *73*(3), 221-248.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, *44*(11), S182-S188.

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms.

Hamre, B. K., Pianta, R. C., Burchinal, M. R., Field, S., LoCasale-Crouch, J., Downer, J. T., Howes, C., LaParo, K., & Scott-Little, C. (2012). A course on effective teacher-child interactions: Effects on teacher beliefs, knowledge, and observed practice. *American Educational Research Journal*, *49*(1), 88-123.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). Building a science of classrooms: Application of the CLASS framework in over 4,000 US early childhood and elementary classrooms. *Downloaded on March*, *27*, 2013.

Harms, T., & Clifford, R. M. (1980). Early Childhood Environment Rating Scale (ECERS).

Harms, T., Clifford, R. M., & Cryer, D. (1998). Early Childhood Environment Rating Scale – Revised (ECERS-R) (Revised ed.).

Harrell-Williams, L. M., & Wolfe, E. W. (2013). The Influence of Between-Dimension Correlation, Misfit, and Test Length on Multidimensional Rasch Model Information-Based Fit Index Accuracy. *Educational and Psychological Measurement*, *73*(4), 672-689.

Harter, S. (1996). Teacher and classmate influences on scholastic motivation, self-esteem, and level of voice in adolescents. In Juvonen, Jaana (Ed); Wentzel, Kathryn R. (Ed). (1996). *Social motivation: Understanding children's school adjustment*., (pp. 11-42). New York, NY, US: Cambridge University Press, xv, 375 pp. Head Start (2003). Head Start child outcomes: Setting the context for the National Reporting System (National Head Start Training and Technical Assistance Resource Center Publication No. 76). *Washington, DC: Head Start Bureau*.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: issues and*

practice, 10(2), 33-41.

Hasan, A., Brinkman, S. A., Jung, H., Kinnell, A., Nakajima, N., & Pradhan, M. P. (2016). The Role of Preschool Quality in Promoting Child Development: Evidence from Rural Indonesia.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900-1902.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, *94*(1), 114-128.

Hestenes, L. L., Kintner-Duffy, V., Wang, Y. C., La Paro, K., Mims, S. U., Crosby, D. & Cassidy, D. J. (2015). Comparisons among quality measures in child care settings: Understanding the use of multiple measures in North Carolina's QRIS and their links to social-emotional development in preschool children. *Early Childhood Research Quarterly*, *30*, 199-214.

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., Humez, A. & Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment, 17*(2-3), 88-106.

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

Howes, C., Burchinal, M. R., 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*(1), 27-50.

Howes, C., & Hamilton, C. E. (1992). Children's relationships with child care teachers: Stability

and concordance with parental attachments. *Child Development, 63*(4), 867-878.

Howes, C., Whitebook, M., & Phillips, D. (1992). Teacher characteristics and effective teaching in child care: Findings from the National Child Care Staffing Study. Child & Youth Care Forum. Special Issue: Meeting the child care needs of the 1990s: Perspectives on day care: II, 21, 399 – 414.

Hubbs-Tait, L., Culp, A. M., Huey, E., Culp, R., Starost, H. J., & Hare, C. (2002). Relation of Head Start attendance to children's cognitive and social outcomes: Moderation by family risk. *Early Childhood Research Quarterly*, *17*(4), 539-558.

Huang, G., Salvucci, S., Peng, S., & Owings, J. (1996). National education longitudinal study of 1988 (NELS: 88) Research framework and issues. *National Center for Education Statistics Working Paper*, (96-03).

Hustedt, J. T., Barnett, W. S., Jung, K., & Goetze, L. D. (2009). The New Mexico PreK Evaluation: Results from the Initial Four Years of a New State Preschool Initiative Final Report.

Hustedt, J. T., Barnett, W. S., Jung, K., & Thomas, J. (2007). The effects of the Arkansas Better Chance Program on young children's school readiness. *National Institute for Early Education Research*.

Hyun, E. (2003). What does the No Child Left Behind Act mean to early childhood teacher educators?: A call for a collective professional rejoinder. *Early Childhood Education Journal*, *31*(2), 119-125.

Janssen, R., & De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research*, *34*(2), 245-268.

Kagan, S. L., Moore, E., & Bredekamp, S. (Eds.). (1998). *Reconsidering Children's Early Development and Learning Toward Common Views and Vocabulary: National Education Goals Panel*. DIANE Publishing.

Kallemeyn, L. M., & DeStefano, L. (2009). The (limited) use of local-level assessment system: A case study of the Head Start National Reporting System and on-going child assessments in a local program. *Early Childhood Research Quarterly*, *24*(2), 157-174.

Kane, M. (2006). Content-related validity evidence in test development. Handbook of test development, 131-153.

Kane, T., Kerr, K., & Pianta, R. (2014). Designing teacher evaluation systems: New guidance from the measures of effective teaching project. John Wiley & Sons.

Katz, L. (1993). Multiple perspectives on the quality of early child- hood programs. ERIC # ED355 041.

Keys, T. D., Farkas, G., Burchinal, M. R., Duncan, G. J., Vandell, D. L., Li, W., Ruzek, E. & Howes, C. (2013). Preschool center quality and school readiness: Quality effects and variation by demographic and child characteristics. *Child Development*, *84*(4), 1171-1190.

Kieffer, K. M. (1998). Why generalizability theory is essential and classical test theory is often inadequate.

Kiefer, T., Robitzsch, A., & Wu, M. (2015). TAM: Test analysis modules (R package version 1.1-0).

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

Kontos, S., Howes, C., Shinn, M., and Galinsky, E. (1995). Quality in Family Child Care and Relative Care, Teachers College Press, New York.

Kontos, S., & Keyes, L. (1999). An ecobehavioral analysis of early childhood classrooms. *Early Childhood Research Quarterly*, *14*(1), 35-50.

Kuhn, T. S., & Hawkins, D. (1963). The structure of scientific revolutions. *American Journal of Physics*, *31*(7), 554-555.

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), n1.

Ladd, G. W., Birch, S. H., & Buhs, E. S. (1999). Children's social and scholastic lives in kindergarten: Related spheres of influence?. *Child Development*, *70*(6), 1373-1400.

Lahti, M., Elicker, J., Zellman, G., & Fiene, R. (2015). Approaches to validating child care quality rating and improvement systems (QRIS): Results from two states with similar QRIS type designs. *Early Childhood Research Quarterly*, *30*, 280-290.

Lambert, M. C., Williams, S. G., Morrison, J. W., Samms-Vaughan, M. E., Mayfield, W. A., & Thornburg, K. R. (2008). Are the indicators for the Language and Reasoning Subscale of the Early Childhood Environment Rating Scale-Revised psychometrically appropriate for Caribbean classrooms?. International Journal of Early Years Education, 16(1), 41-60.

Layzer, J. I., & Goodson, B. D. (2006). The “quality” of early care and education settings Definitional and measurement issues. *Evaluation Review, 30*(5), 556-576.

Le, V. N., Schaack, D. D., & Setodji, C. M. (2015). Identifying baseline and ceiling thresholds

within the Qualistar early learning quality rating and improvement system. *Early Childhood Research Quarterly*, *30*, 215-226.

Leow, C., & Wen, X. (2016). Is Full Day Better Than Half Day? A Propensity Score Analysis of the Association Between Head Start Program Intensity and Children's School Performance in Kindergarten. *Early Education and Development*, 1-16.

Leventhal, T., Selner-O'Hagan, M. B., Brooks-Gunn, J., Bingenheimer, J. B., & Earls, F. J. (2004). The Homelife Interview from the Project on Human Development in Chicago Neighborhoods: Assessment of parenting and home environment for 3-to 15-year-olds. *Parenting*, *4*(2-3), 211-241.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*, 103-122.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85-106.

Linacre, J. M. (2010). Predicting responses from Rasch measures. *Journal of Applied Measurement*, *11*(1), 1.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*.

Ludlow, L. H., & Haley, S. M. (1995). Rasch model logits: Interpretation, use, and transformation. *Educational and Psychological Measurement*, *55*(6), 967-975.

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*(6), 921-943.

Malmberg, L., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality

during teacher education and two years of professional practice. *Journal of Educational Psychology, 102*(4), 916-932.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(3), 200-15.

Mashburn, A. J. (2017). Evaluating the Validity of Classroom Observations in the Head Start Designation Renewal System. *Educational Psychologist, 52*(1), 38-49.

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*(3), 732-749.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

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.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods*, *8*(1), 72.

McCawley, P. (2001). The Logic Model for Program Planning and Evaluation.

McDonald, D. (2009). Elevating the field: using NAEYC early childhood program accreditation to support and reach higher quality in early childhood programs. *Washington: NAEYC*.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

Munton, A. G., Rowland, L., Mooney, A., & Lera, M. J. (1997). Using the Early Childhood

Environment Rating Scale (ECERS) to evaluate quality of nursery provision in England: Some data concerning reliability. *Educational Research*, *39*(1), 99-104.

Najarian, M., Snow, K., Lennon, J., Kinsey, S., & Mulligan, G. (2007). Early childhood longitudinal study, birth cohort (ECLS-B).

National Center on Child Care Quality Improvement (2013). Use of ERS and Other Program Assessment Tools in QRIS. Retrieved from https://qrisguide.acf.hhs.gov/files/QRIS_Program_Assess.pdf

National Center on Education, & the Economy (US). New Commission on the Skills of the American Workforce. (2008). *Tough Choices or Tough Times: The Report of the New Commission on the Skills of the American Workforce*. John Wiley & Sons.

National Center for Health Statistics (US), & National Center for Health Services Research. (2001). *Health, United States*. US Department of Health, Education, and Welfare, Public Health Service, Health Resources Administration, National Center for Health Statistics.

Nelson, G., Westhues, A., & MacLeod, J. (2003). A meta-analysis of longitudinal research on preschool prevention programs for children. *Prevention & Treatment*, *6*(1), 31a.

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

NICHD Early Child Care Research Network. (1999). Child outcomes when child care center classes meet recommended standards for quality. *American Journal of Public Health, 89*, 1072-1077.

NICHD Early Child Care Research Network. (2001). Child care and children's peer interaction at 24 and 36 months: The NICHD study of early child care. *Child Development*, 1478-1500.

OECD. (2016). Education at a Glance 2016 Report.

NICHD Early Child Care Research Network. (2002). Child- care Structure, Process, Outcome: Direct and in- direct effects of child-care quality on young children's development. *Psychological Science, 13*, 199 – 206.

No Child Left Behind (2002). Act of 2001, Pub. L. No. 107-110, § 115. *Stat*, *1425*, 107-110.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C.D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D.P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Levy Paluck, E., Simonsohn, U., Soderberg, C., Spellman, B.A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers E.J., Wilson, R., Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615-631.

Pakarinen, E., Lerkkanen, M., Poikkeus, A., Kiuru, N., Siekkinen, M., Rasku-Puttonen H., & Nurmi, J. (2010): A Validation of the Classroom Assessment Scoring System in Finnish Kindergartens, *Early Education & Development, 21*(1), 95-124.

Parker, E., Atchison, B., & Workman, E. (2016). State Pre-K Funding for 2015-16 Fiscal Year: National Trends in State Preschool Funding. 50-State Review. *Education Commission of the States*.

Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System:

Findings from the prekindergarten year. *The Elementary School Journal, 104*(5), 409-426.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S.L., Yazejian, N., Byler, P., Rustici, J., and Zelazo, J. (2000). The Children of the Cost, Quality, and Outcomes Study Go to School: Technical Report, University of North Carolina at Chapel Hill, Frank Porter Graham Development Center, Chapel Hill.

Perlman, M., Zellman, G. L., & Le, V. N. (2004). Examining the psychometric properties of the early childhood environment rating scale-revised (ECERS-R). *Early Childhood Research Quarterly*, *19*(3), 398-412.

Phillips, D. A., & Howes, C. (1987). Indicators of quality in child care: Review of research. *Quality in child care: What does research tell us*, *1*, 1-20.

Phillips, D., Howes, C., & Whitebook, M. (1991). Child care as an adult work environment. *Journal of Social Issues, 47*(2), 49-70.

Piaget, J. (1952). *The origins of intelligence in children* (Vol. 8, No. 5, pp. 18-1952). New York: International Universities Press.

Pianta, R. C. (2012). Implementing Observation Protocols: Lessons for K-12 Education from the Field of Early Childhood. *Center for American Progress*.

Pianta, R. C., DeCoster, J., Cabell, S., Burchinal, M., Hamre, B. K., Downer, J., LoCasale-Crouch, Wilford, A., & Howes, C. (2014). Dose–response relations between preschool teachers' exposure to components of professional development and increases in quality of their interactions with children. *Early Childhood Research Quarterly, 29*(4), 499-508.

Pianta, R., Downer, J., & Hamre, B. (2016). Quality in Early Education Classrooms: Definitions, Gaps, and Systems. *The Future of Children, 26*(2), 119-137.

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

Pianta, R. C., Steinberg, M. S., & Rollins, K. B. (1995). The first two years of school: Teacher-child relationships and deflections in children's classroom adjustment. *Development and Psychopathology*, *7*(02), 295-312.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. Understanding statistics: *Statistical issues in psychology, education, and the social sciences, 2*(1), 13-43.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raftery, A. E. (1986). Choosing models for cross-classifications. *American Sociological Review*, *51*(1), 145-146.

Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. *Copenhagen: Danish Institute for Educational Research*.

Reckase, M. (2009). *Multidimensional item response theory* (Vol. 150). New York: Springer.

Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2001). Long-term effects of an early childhood intervention on educational achievement and juvenile arrest: A 15-year follow-up of low-income children in public schools. *Jama*, *285*(18), 2339-2346.

Rimm-Kaufman, S. E., Early, D. M., Cox, M. J., Saluja, G., Pianta, R. C., Bradley, R. H., &

Payne, C. (2002). Early behavioral attributes and teachers' sensitivity as predictors of competent behavior in the kindergarten classroom. *Journal of Applied Developmental Psychology*, *23*(4), 451-470.

Ripley, B. D. (2004). Selecting amongst large classes of models. *Methods and models in statistics: In honor of Professor John Nelder, FRS*, 155-170.

Roeser, R. W., Eccles, J. S., & Sameroff, A. J. (2000). School as a context of early adolescents' academic and social-emotional development: A summary of research findings. *The elementary school journal*, 443-471.

Rubin, K. H., Coplan, R. J., Fox, N. A., & Calkins, S. D. (1995). Emotionality, emotion regulation, and preschoolers' social adaptation. *Development and Psychopathology*, *7*(01), 49-62.

Ryan, R. M., Stiller, J. D., & Lynch, J. H. (1994). Representations of relationships to teachers, parents, and friends as predictors of academic motivation and self-esteem. *The Journal of Early Adolescence*, *14*(2), 226-249.

Sabol, T. J., & Pianta, R. C. (2015). Validating Virginia's quality rating and improvement system among state-funded pre-kindergarten programs. *Early Childhood Research Quarterly*, *30*, 183-198.

Sall, S. P. (2014). Maternal Labor Supply and the Availability of Public Pre-K: Evidence from the Introduction of Prekindergarten into American Public Schools. *Economic Inquiry*, *52*(1), 17-34.

Sakai, L. M., Whitebook, M., Wishard, A., & Howes, C. (2004). Evaluating the Early Childhood Environment Rating Scale (ECERS): Assessing differences between the first and revised edition. *Early Childhood Research Quarterly*, *18*(4), 427-445.

Sammons, P., Sylva, K., Hall, J., Siraj, I., Melhuish, E., Taggart, B., & Mathers, S. (2017). Establishing the Effects of Quality in Early Childhood: Comparing evidence from England.

Sandilos, L. E., & DiPerna, J. C. (2013). A Review of Empirical Evidence and Practical Considerations for Early Childhood Classroom Observation Scales. *NHSA Dialog*, *17*(2).

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

Schilder, D., Iruka, I., Dichter, H., & Mathias, D. (2015). Quality Rating and Improvement Systems: Stakeholder Theories of Change and Models of Practice. BUILD Initiative.

Schultz, T. (2015). Early Childhood Governance and Accountability. *Early Childhood Governance: Choices and Consequences*, 95.

Schumacker, R. E., & Smith, E. V. (2007). A Rasch perspective. *Educational and Psychological Measurement*, *67*(3), 394-409.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Scott-Little, C. (2012). A course on effective teacher-child interactions: Effects on teacher beliefs, knowledge, and observed practice. *American Educational Research Journal, 49*(1), 88-123.

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A practical guide to calculating Cohen's f2, a measure of local effect size, from PROC MIXED. *Frontiers in Psychology*, *3*, 111.

Shonkoff, J. P., & Phillips, D. A. (Eds.). (2000). *From neurons to neighborhoods: The science of early childhood development*. National Academies Press.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107.

Silver, R. B., Measelle, J. R., Armstrong, J. M., & Essex, M. J. (2005). Trajectories of classroom externalizing behavior: Contributions of child characteristics, family characteristics, and the teacher–child relationship during the school transition. *Journal of School Psychology*, *43*(1), 39-60.

Singh, J. (2004). Tackling measurement problems with Item Response Theory: Principles, characteristics, and assessment, with an illustrative example. *Journal of Business Research, 57*(2), 184-208.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, *8*(1), 33.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, *51*(3), 541-565.

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

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. Routledge.

Stevens, S. S. (1946). On the theory of scales of measurement.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics.

Stipek, D. (2006). No child left behind comes to preschool. *The Elementary School Journal*, *106*(5), 455-466.

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

The Every Student Succeeds Act (2015). S. 1177. In 114th Congress.

Thompson, J. A., & Sonnenschein, S. (2016). Full-day kindergarten and children's later reading: The role of early word reading. *Journal of Applied Developmental Psychology*, *42*, 58-70.

Tobin, R. M., & Graziano, W. G. (2006). Development of regulatory processes through adolescence: A review of recent empirical studies. *Handbook of Personality Development*, 263-283.

Tout, K., Chien, N., Rothenberg, L., & Li, W. (2014). Implications of QRIS Design for the Distribution of Program Ratings and Linkages between Ratings and Observed Quality. OPRE Research Brief 2014-33. *Administration for Children & Families*.

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., & Boller, K. (2010). *Compendium of quality rating systems and evaluations*. Mathematica Policy Research.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, *16*(4), 8-14.

U.S. Department of Education, Office of Elementary and Secondary Education, Non-Regulatory Guidance Early Learning in the Every Student Succeeds Act: Expanding Opportunities to Support our Youngest Learners, Washington, D.C., 2016

Vandell, D., & Wolfe, B. (2000). *Child care quality: Does it matter and does it need to be improved?* (Vol. 78). University of Wisconsin--Madison, Institute for Research on Poverty.

Virtanen, T. E., Pakarinen, E., Lerkkanen, M. K., Poikkeus, A. M., Siekkinen, M., & Nurmi, J. E. (2017). A Validation Study of Classroom Assessment Scoring System–Secondary in the Finnish School Context. The Journal of Early Adolescence, 0272431617699944.

Vu, J. A., Jeon, H. J., & Howes, C. (2008). Formal education, credential, or both: Early

childhood program classroom practices. *Early Education and Development*, *19*(3), 479-504.

Vygotsky, L. S. (1979). Consciousness as a problem in the psychology of behavior. *Soviet Psychology*, *17*(4), 3-35.

Vygotsky, L. (1991). S. (1978). *Mind in society: The development of higher psychological processes*.

Wang, W. C. (2007). Assessment of differential item functioning. *Journal of Applied Measurement*, *9*(4), 387-408.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, *84*(6), 2112-2130.

Weisner, T. S. (2002). Ecocultural understanding of children's developmental pathways. *Human Development*, *45*(4), 275-281.

Wells, M. B. (2017). Is all support equal?: Head Start preschool teachers' psychological job attitudes. *Teaching and Teacher Education, 63*, 103-115.

Wentzel, K. R. (1998). Social relationships and motivation in middle school: The role of parents, teachers, and peers. *Journal of Educational Psychology*, *90*(2), 202.

West, J., & Rathburn, A. (2004). ECLS-K technical issues. *Presentation at the AERA Institute on Statistical Analysis for Education Policy, San Diego, CA*.

Whitebook, M., & Sakai, L. (2003). Turnover begets turnover: An examination of job and occupational instability among child care center staff. *Early Childhood Research Quarterly, 18*(3), 273-293.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation

of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, *27*(1), 122-154.

WK Kellogg Foundation. (2004). *WK Kellogg Foundation Logic Model Development Guide*. WK Kellogg Foundation.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 3*(1), 3-24.

Wright, B.D., & Linacre, J.M. (2004). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis. Rasch Measurement*. MESA Press, 5835 S. Kimbark Avenue, Chicago, IL 60637.

Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Barata, M., Weiland, C., Gomez, C., Moreno, L., Rolla, A., D'Sha, N., & Arbour, M. C. (2015). Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. *Developmental Psychology, 51*(3), 309.

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W. T. & Zaslow, M. J. (2013). Investing in our future: The evidence base on preschool education. *Ann Arbor, MI: Society for Research in Child Development*.

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

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement*, *36*(1), 1-28.

Appendix A

Table A1. *Descriptive Statistics for the ECERS-R Items*

Item	Mean	SD
1 Indoor Space	5.2	1.8
2 Furnishings for routine care	6.3	1.26
3 Furnishings for relaxation	4.29	1.75
4 Room Arrangement	5.37	1.87
5 Space for Privacy	4.27	1.78
6 Display for Children	4.26	1.54
7 Gross Motor Space	3.93	1.92
8 Gross Motor Equipment	4.45	2.14
9 Greeting/Departing	5.9	1.74
10 Meals/Snacks	3.19	2.36
11 Nap	2.04	2.56
12 Diapering/Toileting	3.5	2.44
13 Health Practice	3.67	2.21
14 Safety Practice	3.88	2.54
15 Books and Pictures	4.4	1.54
16 Encouraging to Communicate	5.6	1.59
17 Language to develop reasoning	4.52	1.84
18 Informal use of language	5.38	1.7
19 Fine motor activities	4.77	1.66
20 Art	4.1	1.62
21 Music and Movement	3.63	1.56
22 Blocks	4.13	1.52
23 Sand and water play	4.04	1.87
24 Dramatic Play	3.92	1.34
25 Nature and Science	3.34	1.74
26 Math	4.1	1.54
27 Use of TV, video or computer	2.52	2.65
28 Promoting acceptance of diversity	4.12	1.61
29 Gross motor supervision	4.89	1.8
30 General supervision	5.43	1.91
31 Discipline	5.37	1.72
32 Staff-child interactions	6.03	1.77
33 Interactions among children	5.78	1.63
34 Schedule of daily play	4.33	2.08
35 Free play	5.06	1.87
36 Group time	5.59	1.85
37 Provisions for exceptional children	1.46	3.44

Note. N = 1400 classrooms, which has been rounded per the ECLS-B user agreement. Descriptive statistics were calculated using the W31C0 weight.

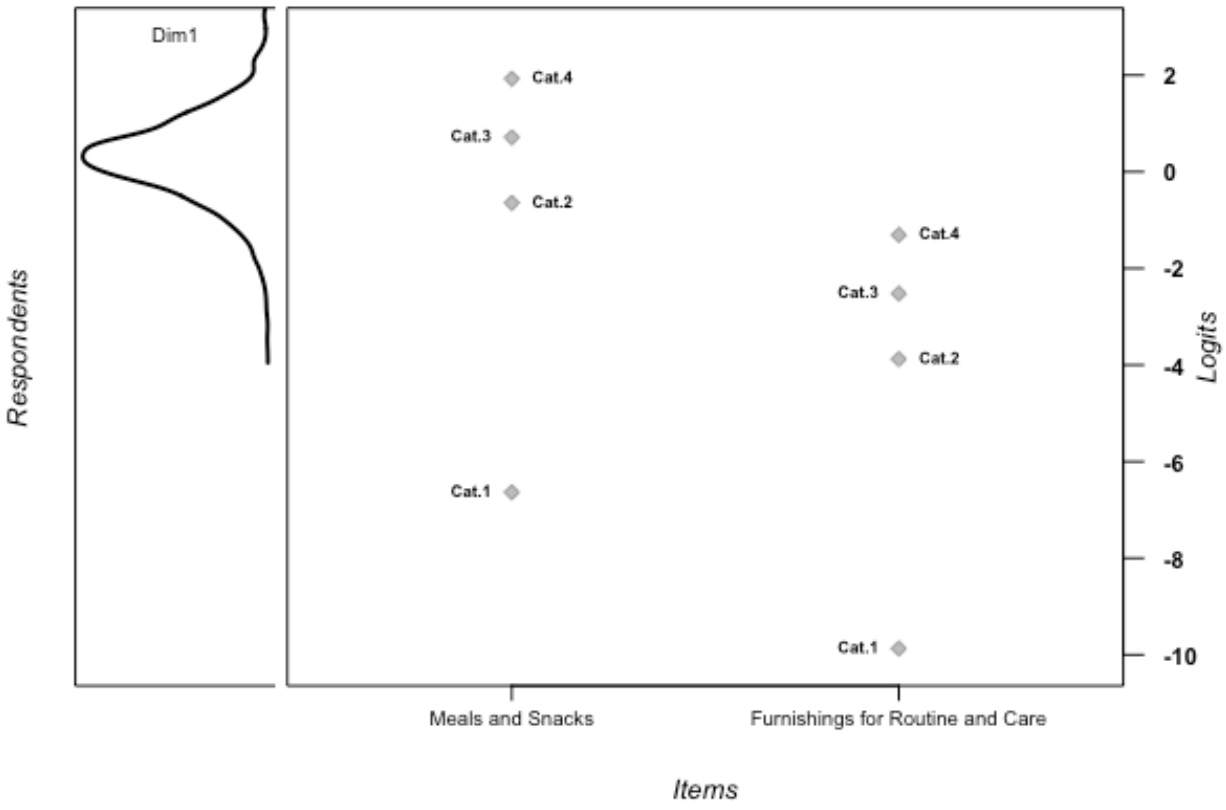

Figure 10. Wright Map for the thresholds for the easiest item (i.e., Furnishings for Routine and Care) and the hardest item (i.e., Meals and Snacks) for the 37-item unidimensional specification of the measure.

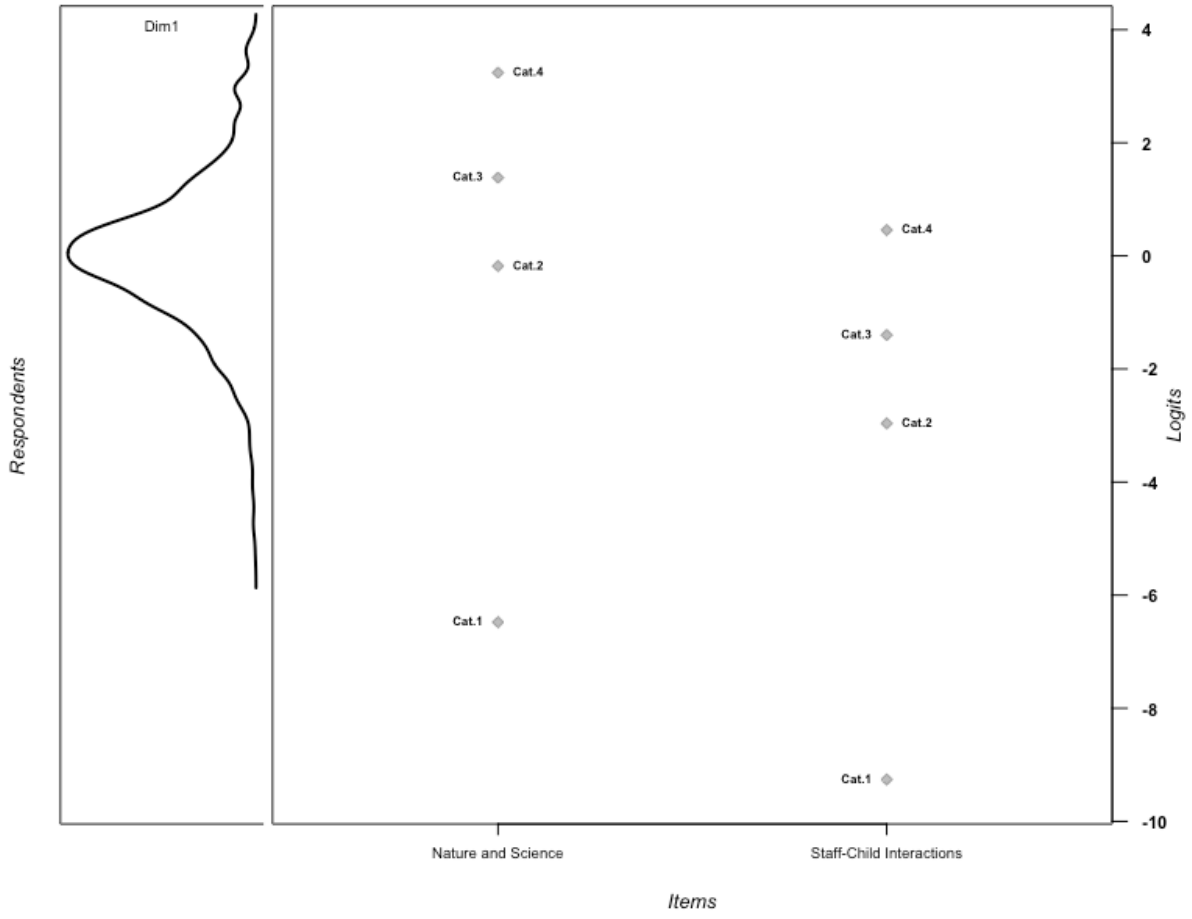

Figure 11. Wright Map for the thresholds for the easiest item (i.e., Staff-Child Interactions) and the hardest item (i.e., Nature and Science) for the 16-item unidimensional specification of the measure.

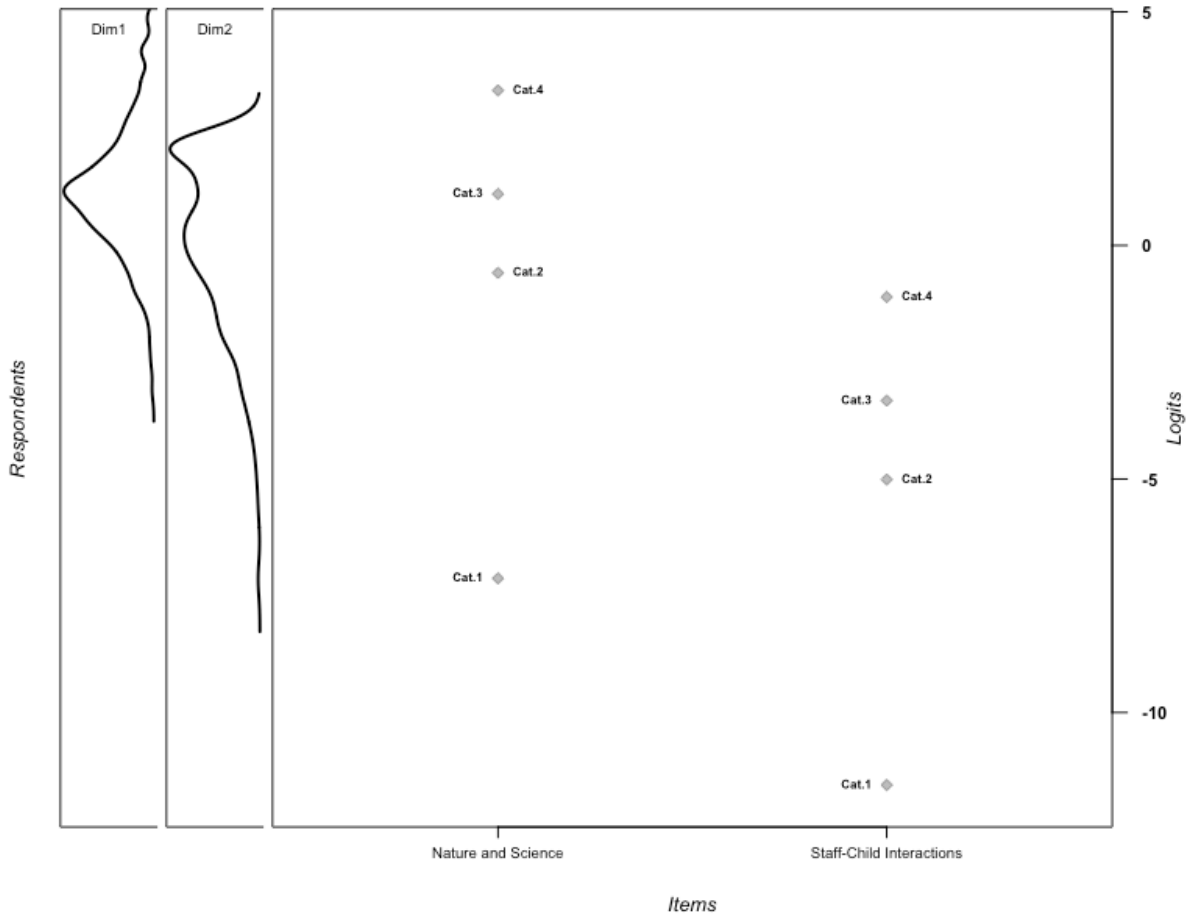

Figure 12. Wright Map for the thresholds for the easiest item (i.e., Staff-Child Interactions) and the hardest item (i.e., Nature and Science) for the Provisions for Learning/Teaching and Interactions specification of the measure. Dim 1 = Provisions for Learning and Dim 2 = Teaching and Interactions.

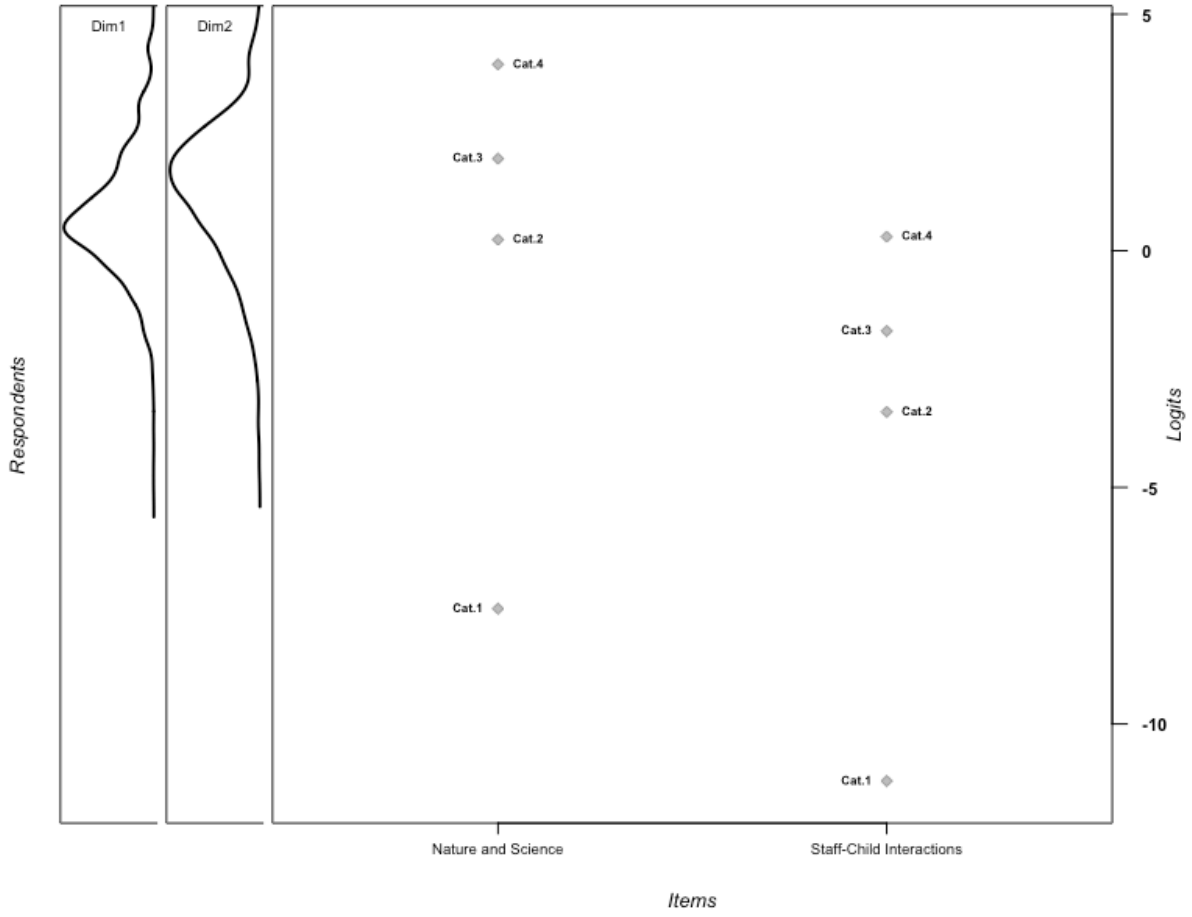

Figure 13. Wright Map for the thresholds for the easiest item (i.e., Staff-Child Interactions) and the hardest item (i.e., Nature and Science) for the Structural/Process specification of the measure. Dim 1 = Structural and Dim 2 = Process.

Table A2. *Differential Item Functioning for Center Type for the 37-Item Unidimensional Model*

	Head Start		Private		State/Local		Other	
Item	DIF	SE	DIF	SE	DIF	SE	DIF	SE
1 Indoor Space	-.04	.06	.32	.07	-.06	.05	-.10	.04
2 Furnishings for routine care	-.39	.08	-.41	.09	.08	.07	.40	.06
3 Furnishings for relaxation	.07	.06	.26	.06	-.31	.05	-.07	.04
4 Room Arrangement	-.41	.06	.42	.07	-.24	.05	.11	.04
5 Space for Privacy	-.05	.06	.06	.06	-.01	.05	.12	.04
6 Display for Children	.01	.06	.05	.06	.00	.05	.11	.04
7 Gross Motor Space	-.06	.06	.34	.06	.11	.04	-.11	.04
8 Gross Motor Equipment	.00	.06	.20	.06	.32	.05	-.32	.04
9 Greeting/Departing	.20	.07	**-1.18**	.08	.46	.06	.32	.05
10 Meals/Snacks	**-.64**	.06	.76	.06	.05	.05	-.21	.04
11 Nap	-.04	.06	.32	.07	-.03	.05	-.19	.04
12 Diapering/Toileting	-.16	.06	.27	.06	.14	.04	-.09	.04
13 Health Practice	-.22	.06	.46	.06	.04	.04	-.19	.04
14 Safety Practice	.04	.06	.24	.06	-.11	.04	-.15	.04
15 Books and Pictures	.26	.06	.07	.06	-.13	.05	-.02	.04
16 Encouraging to Communicate	.06	.07	-.30	.07	-.19	.06	.19	.04
17 Language to develop reasoning	.25	.06	-.07	.06	-.13	.05	-.06	.04
18 Informal use of language	.35	.06	**-.62**	.07	.09	.05	.08	.04
19 Fine motor activities	.05	.06	.11	.07	-.07	.05	.10	.04
20 Art	-.02	.06	.31	.06	-.07	.04	-.08	.04
21 Music and Movement	-.13	.06	.27	.06	.02	.04	-.03	.04
22 Blocks	-.16	.06	.49	.06	-.11	.05	-.03	.04
23 Sand and water play	-.31	.06	.37	.06	.03	.04	-.01	.04
24 Dramatic Play	.10	.06	.11	.06	-.07	.04	-.04	.04
25 Nature and Science	.15	.05	-.18	.06	.04	.04	.07	.04
26 Math	-.04	.06	-.08	.06	-.02	.05	.21	.04
27 Use of TV, video or computer	.04	.06	.02	.07	-.03	.05	-.16	.04
28 Promoting acceptance of diversity	-.26	.06	.32	.06	-.12	.04	.06	.04
29 Gross motor supervision	-.09	.06	.20	.06	-.05	.05	-.17	.04
30 General supervision	.25	.06	-.21	.07	-.05	.05	-.11	.04
31 Discipline	.45	.06	**-.85**	.07	-.04	.05	.14	.04
32 Staff-child interactions	.40	.07	-.02	.08	-.26	.06	-.19	.05
33 Interactions among children	.45	.07	-.49	.08	.00	.06	-.04	.05
34 Schedule of daily play	-.14	.06	-.33	.06	.35	.05	.12	.04
35 Free play	.07	.06	**-.82**	.07	**.54**	.05	.14	.04
36 Group time	-.10	.07	-.49	.07	.28	.05	.22	.04
37 Provisions for exceptional children								

Note. N = 1400, and is rounded to the nearest 50 per ECLS-B data reporting restrictions.

Table A3. *Differential Item Functioning for Teacher Education for the 37-Item Unidimensional Model*

	Less Than High School		High School		Vocational		Some College		Associates		Bachelor		Graduate	
Item	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE
1 Indoor Space	.00	.21	-.26	.13	.24	.23	.23	.08	.19	.08	.20	.06	.15	.08
2 Furnishings for routine care	**1.38**	.23	.26	.17	.49	.32	.47	.11	.31	.11	.42	.08	.28	.11
3 Furnishings for relaxation	-.28	.20	-.16	.11	.11	.22	-.21	.07	-.12	.07	-.12	.05	-.09	.07
4 Room Arrangement	.38	.19	-.15	.12	.14	.22	-.05	.08	-.39	.08	-.22	.06	-.19	.08
5 Space for Privacy	-.11	.20	.10	.11	**-.72**	.23	-.17	.07	-.10	.07	-.14	.05	-.21	.07
6 Display for Children	.17	.20	-.16	.11	.12	.21	-.12	.07	-.06	.07	-.07	.05	.01	.07
7 Gross Motor Space	-.33	.19	-.33	.11	**-.57**	.22	.17	.07	.31	.07	.27	.05	**.56**	.07
8 Gross Motor Equipment	-.33	.21	-.24	.12	.16	.23	.09	.08	.18	.07	.37	.05	**.92**	.07
9 Greeting/Departing	.06	.22	-.04	.14	-.39	.28	.10	.09	.17	.09	.21	.07	.06	.09
10 Meals/Snacks	-.41	.21	-.21	.12	**-.53**	.22	.07	.08	.00	.07	.06	.05	.37	.07
11 Nap	-.06	.21	-.14	.12	-.33	.23	.05	.08	.16	.08	.20	.06	.32	.08
12 Diapering/Toileting	.24	.21	-.28	.11	.18	.21	-.06	.07	.09	.07	.09	.05	.16	.07
13 Health Practice	.11	.20	-.21	.11	.34	.21	.09	.07	.06	.07	.13	.05	.21	.07
14 Safety Practice	-.24	.19	-.24	.11	-.46	.22	.25	.07	.42	.07	.23	.05	.37	.07
15 Books and Pictures	.04	.19	.21	.11	.10	.22	-.03	.08	.05	.07	-.07	.05	-.06	.07
16 Encouraging to Communicate	.62	.20	.29	.12	-.12	.25	-.03	.09	-.05	.08	-.18	.07	**-.51**	.09
17 Language to develop reasoning	.03	.19	.26	.11	.13	.21	.12	.07	-.01	.07	-.04	.05	-.45	.07
18 Informal use of language	-.17	.21	.26	.12	.29	.23	.12	.08	.04	.08	-.02	.06	-.28	.09
19 Fine motor activities	.19	.19	.08	.11	-.13	.22	-.08	.08	-.11	.07	-.08	.06	-.16	.07
20 Art	.02	.20	.05	.11	.31	.21	-.07	.07	-.09	.07	.02	.05	-.01	.07
21 Music and Movement	-.20	.20	-.14	.11	.04	.21	-.14	.07	.03	.07	.06	.05	.15	.07
22 Blocks	-.22	.20	-.18	.11	.04	.22	-.22	.08	-.12	.07	-.17	.05	-.08	.07
23 Sand and water play	-.01	.20	-.18	.11	-.25	.22	-.23	.07	-.16	.07	-.14	.05	-.23	.07
24 Dramatic Play	.17	.20	-.04	.11	-.35	.22	-.03	.07	.16	.07	.14	.05	.36	.07
25 Nature and Science	-.19	.21	-.03	.12	-.02	.23	-.38	.07	-.31	.07	-.26	.05	-.27	.07
26 Math	.17	.20	.11	.11	.02	.21	-.10	.07	-.15	.07	-.16	.05	-.26	.07
27 Use of TV, video or computer	.10	.23	-.14	.12	.17	.24	.12	.08	.05	.07	.05	.06	-.07	.07
28 Promoting acceptance of diversity	-.16	.19	.24	.11	.05	.21	-.03	.07	-.05	.07	.17	.05	.28	.07
29 Gross motor supervision	-.14	.20	.08	.12	.31	.23	.32	.08	.36	.07	.15	.06	-.02	.08
30 General supervision	.08	.20	.08	.12	.24	.23	.24	.08	-.13	.08	-.14	.06	-.39	.08
31 Discipline	.16	.20	-.09	.12	.24	.22	.02	.08	-.10	.08	-.32	.06	-.55	.09
32 Staff-child interactions	.34	.21	-.15	.13	.06	.25	.01	.09	-.17	.09	-.27	.07	-.37	.09
33 Interactions among children	.03	.22	.34	.13	.31	.25	-.04	.09	.05	.09	-.13	.07	-.30	.09

Table A3. *Differential Item Functioning for Teacher Education for the 37-Item Unidimensional Model*

	Less Than High School		High School		Vocational		Some College		Associates		Bachelor		Graduate	
Item	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE
34 Schedule of daily play	-.40	.20	-.22	.11	-.13	.22	-.09	.08	-.12	.07	.00	.05	.25	.07
	Less Than High School		High School		Vocational		Some College		Associates		Bachelor		Graduate	
Item	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE
35 Free play	-.01	.20	.09	.12	-.46	.24	-.10	.08	-.21	.08	.06	.06	.12	.07
36 Group time	.17	.20	.36	.12	.06	.24	-.20	.09	-.17	.08	-.09	.06	.08	.08
37 Provisions for exceptional children														

Note. N = 1400, and is rounded to the nearest 50 per ECLS-B data reporting restrictions.

Table A4. *Differential Item Functioning for Half Time Program Status for the 37-Item Unidimensional Model*

Item	DIF	SE
1 Indoor Space	.06	.03
2 Furnishings for routine care	.01	.05
3 Furnishings for relaxation	-.06	.03
4 Room Arrangement	-.07	.03
5 Space for Privacy	-.08	.03
6 Display for Children	-.07	.03
7 Gross Motor Space	.02	.03
8 Gross Motor Equipment	.05	.03
9 Greeting/Departing	-.03	.04
10 Meals/Snacks	.08	.03
11 Nap	.02	.04
12 Diapering/Toileting	.04	.03
13 Health Practice	-.03	.03
14 Safety Practice	.04	.03
15 Books and Pictures	-.06	.03
16 Encouraging to Communicate	-.04	.04
17 Language to develop reasoning	.04	.03
18 Informal use of language	.16	.03
19 Fine motor activities	-.07	.03
20 Art	.02	.03
21 Music and Movement	-.13	.03
22 Blocks	-.05	.03
23 Sand and water play	-.09	.03
24 Dramatic Play	-.02	.03
25 Nature and Science	-.12	.03
26 Math	-.09	.03
27 Use of TV, video or computer	.02	.03
28 Promoting acceptance of diversity	-.06	.03
29 Gross motor supervision	.08	.03
30 General supervision	.10	.04
31 Discipline	.06	.04
32 Staff-child interactions	.23	.04
33 Interactions among children	.13	.04
34 Schedule of daily play	-.10	.03
35 Free play	-.02	.03
36 Group time	-.02	.04
37 Provisions for exceptional children		

Note. N = 1400, and is rounded to the nearest 50 per ECLS-B data reporting restrictions.

Table A5. *Differential Item Functioning for Center Type for the 16-Item Unidimensional Model*

	Head Start		Private		State/Local		Other	
Item	DIF	SE	DIF	SE	DIF	SE	DIF	SE
3 Furnishings for relaxation	-.10	.05	.44	.06	-.32	.04	-.11	.04
5 Space for Privacy	-.25	.05	.18	.06	.06	.04	.13	.04
15 Books and Pictures	.15	.05	.19	.06	-.09	.04	-.04	.04
17 Language to develop reasoning	-.12	.06	.24	.06	-.02	.05	.12	.04
18 Informal use of language	-.22	.05	.50	.06	.00	.04	-.12	.04
19 Fine motor activities	-.40	.05	.74	.06	-.06	.04	-.06	.04
20 Art	-.07	.05	.24	.06	-.02	.04	-.07	.04
22 Blocks	.00	.05	-.14	.06	.13	.04	.06	.03
24 Dramatic Play	-.24	.05	-.01	.06	.06	.04	.25	.04
25 Nature and Science	.13	.05	.01	.06	-.09	.04	-.10	.04
26 Math	.25	.06	-.62	.06	.16	.05	.07	.04
30 General supervision	.13	.06	-.15	.06	.00	.05	-.16	.04
31 Discipline	.36	.06	-.86	.06	-.02	.05	.14	.04
32 Staff-child interactions	.32	.06	.11	.07	-.26	.05	-.25	.04
33 Interactions among children	.36	.06	-.40	.07	.05	.05	-.09	.04
36 Group time								

Note. N = 1400, and is rounded to the nearest 50 per ECLS-B data reporting restrictions.

Table A6. *Differential Item Functioning for Teacher Education for the 16-Item Unidimensional Model*

	Less Than High School		High School		Vocational		Some College		Associates		Bachelor		Graduate	
Item	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE
3 Furnishings for relaxation	-.40	.17	-.30	.10	.10	.20	-.19	.07	-.05	.07	.00	.05	.12	.07
5 Space for Privacy	-.17	.17	.05	.10	-.97	.21	-.10	.07	-.01	.07	-.01	.05	-.01	.07
15 Books and Pictures	.01	.17	.17	.10	.07	.20	.04	.07	.16	.07	.05	.05	.16	.07
17 Language to develop reasoning	.21	.17	.00	.10	-.23	.21	-.02	.07	-.05	.07	.03	.05	.02	.07
18 Informal use of language	-.04	.17	-.04	.10	.35	.20	-.01	.07	-.02	.07	.17	.05	.22	.07
19 Fine motor activities	-.31	.17	-.32	.10	.01	.20	-.20	.07	-.05	.07	-.07	.05	.15	.07
20 Art	.16	.17	-.14	.10	**-.51**	.21	.04	.07	.30	.07	.32	.05	**.71**	.07
22 Blocks	-.31	.17	-.09	.10	-.01	.21	-.34	.07	-.25	.07	-.13	.05	-.04	.07
24 Dramatic Play	.15	.17	.05	.10	-.02	.20	-.04	.07	-.10	.07	-.06	.05	-.09	.07
25 Nature and Science	-.01	.17	.23	.10	.10	.20	.24	.07	.08	.07	.09	.05	-.36	.07
26 Math	-.23	.18	.24	.10	.31	.21	.22	.08	.15	.08	.11	.06	-.13	.08
30 General supervision	.09	.17	-.01	.11	.23	.21	.38	.07	-.07	.08	-.04	.06	-.27	.08
31 Discipline	.18	.17	-.22	.11	.24	.21	.09	.07	-.04	.07	-.27	.06	-.47	.08
32 Staff-child interactions	.41	.18	-.32	.11	.01	.22	.07	.08	-.10	.08	-.20	.06	-.22	.08
33 Interactions among children	.03	.18	.32	.11	.33	.22	.02	.08	.17	.08	-.02	.06	-.13	.08
36 Group time														

Note. N = 1400, and is rounded to the nearest 50 per ECLS-B data reporting restrictions.

Table A7. *Differential Item Functioning for Half Time Status for the 16-Item Unidimensional Model*

Item	DIF	SE
3 Furnishings for relaxation	-.08	.03
5 Space for Privacy	-.11	.03
15 Books and Pictures	-.09	.03
17 Language to develop reasoning	-.10	.03
18 Informal use of language	.01	.03
19 Fine motor activities	-.08	.03
20 Art	-.03	.03
22 Blocks	-.16	.03
24 Dramatic Play	-.13	.03
25 Nature and Science	.04	.03
26 Math	.19	.03
30 General supervision	.11	.03
31 Discipline	.06	.03
32 Staff-child interactions	.27	.04
33 Interactions among children	.15	.03
36 Group time		

Note. N = 1400, and is rounded to the nearest 50 per ECLS-B data reporting restrictions.

Table A8. *Differential Item Functioning for Center Type for the Provisions for Learning/Teaching Interactions 2-Dimension Model*

	Head Start		Private		State/Local		Other	
Item	DIF	SE	DIF	SE	DIF	SE	DIF	SE
3 Furnishings for relaxation	-.03	.05	-.01	.05	-.21	.04	.07	.03
5 Space for Privacy	-.14	.05	-.07	.06	.14	.04	.14	.03
15 Books and Pictures	.26	.06	-.46	.06	.05	.05	.07	.04
17 Language to develop reasoning	-.06	.06	-.05	.06	-.07	.05	.15	.04
18 Informal use of language	-.17	.05	.48	.06	-.09	.04	-.04	.03
19 Fine motor activities	**-.64**	.06	**.92**	.06	-.10	.04	.02	.04
20 Art	-.01	.05	-.07	.06	.09	.04	.03	.04
22 Blocks	.03	.05	-.06	.05	.09	.04	.07	.03
24 Dramatic Play	-.42	.06	.17	.06	-.06	.05	.24	.04
25 Nature and Science	.02	.05	-.04	.06	-.06	.04	-.03	.04
26 Math	.04	.06	-.28	.07	.17	.05	-.07	.04
30 General supervision	.21	.06	-.04	.06	-.02	.05	-.24	.04
31 Discipline	.32	.06	**-.50**	.06	-.09	.05	.02	.04
32 Staff-child interactions	.11	.06	**.60**	.06	-.23	.05	-.36	.04
33 Interactions among children	.38	.06	-.40	.07	.14	.05	-.11	.04
36 Group time								

Note. N = 1400, and is rounded to the nearest 50 per ECLS-B data reporting restrictions.

Table A9. *Differential Item Functioning for Teacher Education for the Provisions for Learning/Teaching Interactions 2-Dimension Model*

	Less Than High School		High School		Vocational		Some College		Associates		Bachelor		Graduate	
Item	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE
3 Furnishings for relaxation	-.12	.14	-.25	.09	.03	.19	-.09	.07	.05	.07	.09	.05	.21	.07
5 Space for Privacy	-.08	.14	.09	.09	**-.67**	.21	-.05	.07	.07	.07	.03	.05	.17	.07
15 Books and Pictures	.25	.15	.14	.09	-.01	.21	.08	.07	.14	.08	-.21	.06	.02	.08
17 Language to develop reasoning	.22	.15	-.05	.09	-.14	.21	.06	.07	.03	.08	-.01	.06	-.07	.08
18 Informal use of language	.05	.15	-.10	.09	.34	.19	.10	.07	.08	.07	.13	.05	.26	.07
19 Fine motor activities	.02	.14	-.27	.09	.16	.19	-.15	.07	-.09	.08	-.05	.05	.02	.07
20 Art	.32	.15	-.16	.09	-.43	.23	.14	.07	.28	.08	.35	.05	**.56**	.07
22 Blocks	-.35	.15	-.06	.08	.06	.18	-.24	.06	-.08	.06	-.04	.04	-.07	.06
24 Dramatic Play	.28	.14	.13	.09	.14	.20	-.15	.07	-.27	.08	-.28	.06	-.42	.08
25 Nature and Science	-.26	.15	.21	.09	.47	.20	.20	.07	-.02	.08	.06	.05	-.44	.08
26 Math	-.10	.18	.07	.10	-.06	.25	-.03	.08	-.13	.10	-.03	.07	.02	.08
30 General supervision	-.15	.16	-.11	.10	.17	.22	.24	.07	-.07	.08	-.01	.06	-.18	.08
31 Discipline	.10	.16	-.02	.10	.14	.23	.06	.08	.03	.09	-.26	.06	-.35	.08
32 Staff-child interactions	.18	.16	-.20	.10	-.09	.24	.13	.08	.03	.09	.06	.06	-.19	.08
33 Interactions among children	-.32	.18	.14	.10	-.08	.25	-.15	.09	.11	.09	.09	.07	-.02	.08
36 Group time														

Note. N = 1400, and is rounded to the nearest 50 per ECLS-B data reporting restrictions.

Table A10. *Differential Item Functioning Half Time Program Status for the Provisions for Learning/Teaching Interactions 2-Dimension Model*

Item	DIF	SE
3 Furnishings for relaxation	-.06	.03
5 Space for Privacy	-.05	.03
15 Books and Pictures	-.09	.03
17 Language to develop reasoning	-.09	.03
18 Informal use of language	.01	.03
19 Fine motor activities	-.08	.03
20 Art	.00	.03
22 Blocks	-.11	.03
24 Dramatic Play	-.14	.03
25 Nature and Science	.01	.03
26 Math	.09	.04
30 General supervision	.10	.03
31 Discipline	.10	.03
32 Staff-child interactions	.27	.04
33 Interactions among children	.04	.04
36 Group time		

Note. N = 1400, and is rounded to the nearest 50 per ECLS-B data reporting restrictions.

Table A11. *Differential Item Functioning for Center Type for the Structural/Process 2-Dimension Within-Item Model*

	Head Start		Private		State/Local		Other	
Item	DIF	SE	DIF	SE	DIF	SE	DIF	SE
3 Furnishings for relaxation	-.03	.03	.04	.03	-.07	.02	-.01	.02
5 Space for Privacy	-.14	.03	.15	.03	-.02	.02	.06	.02
15 Books and Pictures	.08	.03	.20	.03	-.12	.02	-.03	.02
17 Language to develop reasoning	-.05	.03	-.04	.03	.07	.02	.11	.02
18 Informal use of language	-.09	.03	.25	.03	-.02	.02	-.07	.02
19 Fine motor activities	-.16	.03	.20	.03	.04	.02	.01	.02
20 Art	-.01	.03	-.08	.03	.07	.02	.02	.02
22 Blocks	.01	.03	-.07	.03	.04	.02	.02	.02
24 Dramatic Play	-.11	.03	.04	.03	-.01	.02	.14	.02
25 Nature and Science	.06	.03	-.06	.03	-.03	.02	-.05	.02
26 Math	.11	.03	-.22	.03	.05	.02	.01	.02
30 General supervision	.07	.03	-.01	.03	-.02	.02	-.13	.02
31 Discipline	.11	.03	-.22	.03	-.06	.02	.01	.02
32 Staff-child interactions	.17	.03	.10	.03	-.15	.03	-.14	.02
33 Interactions among children	.11	.03	-.16	.03	.08	.02	-.03	.02
36 Group time								

Note. N = 1400, and is rounded to the nearest 50 per ECLS-B data reporting restrictions.

Table A12. *Differential Item Functioning for Teacher Education for the Structural/Process 2-Dimension Within-Item Model*

	Less Than High School		High School		Vocational		Some College		Associates		Bachelor		Graduate	
Item	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE	DIF	SE
3 Furnishings for relaxation	-.17	.09	-.21	.06	.12	.11	-.11	.04	-.05	.03	-.03	.03	.05	.03
5 Space for Privacy	-.13	.10	.13	.06	-.48	.11	-.04	.04	.01	.04	-.03	.03	-.02	.04
15 Books and Pictures	.13	.10	.22	.06	.01	.11	.09	.04	.13	.04	.06	.03	.09	.04
17 Language to develop reasoning	.09	.09	-.06	.06	-.11	.10	-.04	.04	-.07	.03	-.01	.03	.03	.04
18 Informal use of language	-.01	.10	.07	.06	.11	.11	.00	.04	.03	.04	.09	.03	.18	.04
19 Fine motor activities	-.19	.10	-.19	.06	-.03	.11	-.10	.04	-.05	.04	-.04	.03	.08	.04
20 Art	.06	.10	-.15	.06	-.19	.10	.01	.04	.12	.04	.17	.03	.39	.04
22 Blocks	-.14	.10	.03	.06	-.04	.12	-.21	.04	-.18	.04	-.15	.03	-.11	.04
24 Dramatic Play	.04	.10	.12	.06	-.08	.11	.01	.04	-.04	.04	-.05	.03	-.05	.04
25 Nature and Science	-.03	.09	.07	.06	.07	.11	.11	.04	.00	.04	.00	.03	-.21	.04
26 Math	-.05	.09	.08	.06	.13	.11	.14	.04	.07	.04	.09	.03	-.06	.04
30 General supervision	.04	.09	-.03	.06	.18	.11	.22	.04	-.02	.04	.04	.03	-.13	.04
31 Discipline	.04	.09	-.12	.06	.07	.11	.02	.04	.01	.04	-.13	.03	-.25	.04
32 Staff-child interactions	.21	.09	-.24	.06	.04	.11	-.01	.04	-.08	.04	-.13	.03	-.13	.04
33 Interactions among children	.04	.09	.14	.06	.19	.11	.03	.04	.17	.04	.08	.03	.00	.04
36 Group time														

Note. N = 1400, and is rounded to the nearest 50 per ECLS-B data reporting restrictions.

Table A13. *Differential Item Functioning for Half Time Program Status for the Structural/Process 2-Dimension Within Item Model*

Item	DIF	SE
3 Furnishings for relaxation	-.03	.02
5 Space for Privacy	-.07	.02
15 Books and Pictures	-.06	.02
17 Language to develop reasoning	-.04	.02
18 Informal use of language	.01	.02
19 Fine motor activities	-.02	.02
20 Art	-.01	.02
22 Blocks	-.11	.02
24 Dramatic Play	-.07	.02
25 Nature and Science	.02	.02
26 Math	.10	.02
30 General supervision	.07	.02
31 Discipline	.02	.02
32 Staff-child interactions	.14	.02
33 Interactions among children	.07	.02
36 Group time		

Note. N = 1400, and is rounded to the nearest 50 per ECLS-B data reporting restrictions.

Appendix B

Table B1. *Multivariate Regression for Children's Reading Scores for the 37-Item Unidimensional Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Child and Family Covariates												
Intercept	-.21	.17	-1.30	.42	-2.32	1.89	-.22	.17	-1.30	.42	-2.34	1.91
Female	.26	.05	5.75	.11	-.31	.83	.26	.04	6.69	.09	-.24	.76
Black	-.08	.03	-2.37	.25	-.49	.34	-.08	.04	-2.11	.28	-.53	.38
Hispanic	-.33	.10	-3.33	.19	-1.59	.93	-.33	.10	-3.46	.18	-1.55	.88
Asian	-.02	.07	-.29	.82	-.88	.84	-.02	.06	-.36	.78	-.80	.75
Native Hawaiian or Pacific Islander	.04	.28	.13	.92	-3.48	3.55	.04	.28	.14	.91	-3.55	3.63
American Indian	.15	.16	.97	.51	-1.85	2.16	.16	.15	1.02	.49	-1.79	2.10
Not Hispanic	-.22	.04	-5.61	.11	-.72	.28	-.22	.03	-6.83	.09	-.64	.19
Two Races Not Hispanic	.16	.14	1.16	.45	-1.56	1.88	.16	.14	1.14	.46	-1.58	1.89
Receives WIC Benefits	.02	.05	.52	.70	-.56	.61	.02	.05	.55	.68	-.55	.60
Child Age in Months	.06	.00	39.80	.02	.04	.08	.06	.00	40.05	.02	.04	.08
Less Than 35 Hours a Week	-.02	.09	-.19	.88	-1.20	1.16	-.02	.09	-.22	.87	-1.15	1.11
Looking for Work	-.20	.20	-1.01	.50	-2.79	2.38	-.21	.19	-1.07	.48	-2.67	2.26
Not in the Labor Force	.00	.06	.06	.96	-.78	.79	.00	.06	.03	.98	-.70	.71
Mother's Age	.00	.00	4.53	.14	-.01	.02	.00	.00	4.27	.15	-.01	.02
Number of Children in House < 18 Years Old	-.08	.02	-4.36	.14	-.33	.16	-.08	.02	-4.28	.15	-.33	.17
High School	.08	.12	.66	.63	-1.44	1.60	.08	.13	.66	.63	-1.52	1.69
Some College	.11	.15	.76	.59	-1.80	2.03	.12	.17	.73	.60	-2.00	2.25
Bachelor	.12	.05	2.21	.27	-.56	.80	.13	.07	1.73	.33	-.80	1.05
Graduate	.35	.14	2.57	.24	-1.37	2.07	.35	.14	2.44	.25	-1.48	2.18
Family SES Status	.25	.03	7.45	.09	-.18	.68	.25	.03	8.71	.07	-.12	.62
Suburban	-.23	.03	-6.63	.10	-.68	.21	-.23	.04	-6.07	.10	-.72	.25
Urban	-.12	.13	-.91	.53	-1.82	1.57	-.12	.13	-.94	.52	-1.75	1.51
Teacher and Center Covariates												
Black	-.07	.08	-.81	.57	-1.14	1.01	-.07	.08	-.81	.57	-1.14	1.01
Latino	-.17	.07	-2.39	.25	-1.07	.73	-.17	.06	-2.80	.22	-.95	.61
Other	-.14	.06	-2.18	.27	-.96	.68	-.14	.06	-2.39	.25	-.87	.60
High School	-.33	.12	-2.85	.22	-1.82	1.15	-.34	.09	-3.62	.17	-1.53	.85
Vocational	-.36	.16	-2.25	.27	-2.42	1.69	-.37	.13	-2.94	.21	-1.99	1.24
Some College	-.18	.07	-2.66	.23	-1.07	.70	-.19	.05	-3.96	.16	-.81	.42
Associate	-.13	.10	-1.35	.41	-1.34	1.08	-.14	.07	-1.93	.31	-1.04	.77
Bachelor	-.20	.04	-5.47	.12	-.66	.26	-.21	.01	-26.49	.02	-.31	-.11
Graduate	-.29	.06	-4.98	.13	-1.02	.44	-.30	.09	-3.28	.19	-1.45	.86
Private	.25	.16	1.52	.37	-1.83	2.33	.25	.16	1.56	.36	-1.77	2.26
State/Local	-.09	.04	-2.45	.25	-.53	.36	-.08	.05	-1.77	.33	-.67	.50
Other	-.14	.02	-6.31	.10	-.41	.14	-.13	.03	-4.29	.15	-.53	.26
Program Part Time	-.04	.04	-1.06	.48	-.53	.45	-.04	.03	-1.15	.46	-.47	.39
Rasch Scores							.01	.03	.34	.79	-.38	.40
R2	.40						.40					
Cohen's f2	.67						.67					

Note. N = 950, and rounded to the nearest 50 per the ECLS-B user agreement. A priori power analysis using GPower version 3.1 (2009) indicated that a sample size of N=82 would be sufficient to detect an effect for the Rasch score, assuming f^2=.10, α = .05, using, 36 predictors, with power = .80. A post hoc sensitivity analysis revealed that given α = .05, power = .80, N = 977 and 36 predictors, there was enough power to detect an effect size of f^2=.01.

Table B2. *Multivariate Regression for Children's Reading Scores for the 16-Item Unidimensional Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Child and Family Covariates												
Intercept	-.21	.17	-1.30	.42	-2.32	1.89	-.22	.16	-1.36	.40	-2.25	1.82
Female	.26	.05	5.75	.11	-.31	.83	.27	.04	6.69	.09	-.24	.77
Black	-.08	.03	-2.37	.25	-.49	.34	-.07	.04	-1.88	.31	-.55	.41
Hispanic	-.33	.10	-3.33	.19	-1.59	.93	-.33	.10	-3.44	.18	-1.54	.89
Asian	-.02	.07	-.29	.82	-.88	.84	-.02	.07	-.35	.79	-.86	.82
Native Hawaiian or Pacific Islander	.04	.28	.13	.92	-3.48	3.55	.04	.28	.16	.90	-3.54	3.63
Amerian Indian	.15	.16	.97	.51	-1.85	2.16	.15	.16	.89	.54	-1.94	2.24
Not Hispanic	-.22	.04	-5.61	.11	-.72	.28	-.22	.03	-7.42	.09	-.61	.16
Two Races Not Hispanic	.16	.14	1.16	.45	-1.56	1.88	.16	.13	1.17	.45	-1.55	1.86
Receives WIC Benefits	.02	.05	.52	.70	-.56	.61	.02	.05	.51	.70	-.58	.63
Child Age in Months	.06	.00	39.80	.02	.04	.08	.06	.00	39.55	.02	.04	.08
Less Than 35 Hours a Week	-.02	.09	-.19	.88	-1.20	1.16	-.02	.09	-.26	.84	-1.17	1.12
Looking for Work	-.20	.20	-1.01	.50	-2.79	2.38	-.21	.19	-1.14	.46	-2.61	2.18
Not in the Labor Force	.00	.06	.06	.96	-.78	.79	.00	.06	-.06	.96	-.72	.71
Mother's Age	.00	.00	4.53	.14	-.01	.02	.00	.00	4.10	.15	-.01	.02
Number of Children in House < 18												
Years Old	-.08	.02	-4.36	.14	-.33	.16	-.08	.02	-3.90	.16	-.36	.19
High School	.08	.12	.66	.63	-1.44	1.60	.09	.12	.77	.58	-1.42	1.60
Some College	.11	.15	.76	.59	-1.80	2.03	.14	.16	.87	.54	-1.88	2.16
Bachelor	.12	.05	2.21	.27	-.56	.80	.14	.07	2.12	.28	-.69	.97
Graduate	.35	.14	2.57	.24	-1.37	2.07	.36	.14	2.58	.24	-1.42	2.15
Family SES Status	.25	.03	7.45	.09	-.18	.68	.25	.03	8.93	.07	-.11	.60
Suburban	-.23	.03	-6.63	.10	-.68	.21	-.23	.05	-4.89	.13	-.82	.37
Urban	-.12	.13	-.91	.53	-1.82	1.57	-.12	.12	-.95	.52	-1.69	1.45
Teacher and Center Covariates												
Black	-.07	.08	-.81	.57	-1.14	1.01	-.07	.09	-.79	.58	-1.20	1.06
Latino	-.17	.07	-2.39	.25	-1.07	.73	-.17	.06	-2.80	.22	-.97	.62
Other	-.14	.06	-2.18	.27	-.96	.68	-.13	.06	-2.27	.26	-.89	.62
High School	-.33	.12	-2.85	.22	-1.82	1.15	-.35	.10	-3.46	.18	-1.64	.94
Vocational	-.36	.16	-2.25	.27	-2.42	1.69	-.39	.12	-3.34	.19	-1.87	1.09
Some College	-.18	.07	-2.66	.23	-1.07	.70	-.21	.05	-3.79	.16	-.90	.49
Associate	-.13	.10	-1.35	.41	-1.34	1.08	-.15	.08	-1.96	.30	-1.15	.84
Bachelor	-.20	.04	-5.47	.12	-.66	.26	-.23	.01	-16.66	.04	-.40	-.05
Graduate	-.29	.06	-4.98	.13	-1.02	.44	-.32	.09	-3.78	.17	-1.41	.76
Private	.25	.16	1.52	.37	-1.83	2.33	.24	.15	1.57	.36	-1.70	2.18
State/Local	-.09	.04	-2.45	.25	-.53	.36	-.08	.04	-2.12	.28	-.56	.40
Other	-.14	.02	-6.31	.10	-.41	.14	-.13	.03	-5.18	.12	-.45	.19
Program Part Time	-.04	.04	-1.06	.48	-.53	.45	-.04	.04	-1.02	.50	-.48	.41
Rasch Scores							.02	.02	1.26	.43	-.21	.25
R2	.40						.40					
f2	.67						.67					

Table B2. *Multivariate Regression for Children's Reading Scores for the 16-Item Unidimensional Specification of the Measure*

Note. N = 950, and rounded to the nearest 50 per the ECLS-B user agreement. A priori power analysis using GPower version 3.1 (2009) indicated that a sample size of N=82 would be sufficient to detect an effect for the Rasch score, assuming f^2=.10, α = .05, using, 36 predictors, with power = .80. A post hoc sensitivity analysis revealed that given α = .05, power = .80, N = 977 and 36 predictors, there was enough power to detect an effect size of f^2=.01.

Table B3. *Multivariate Regression for Children's Reading Scores for the Provisions for Learning/Teaching Interactions 2-Dimension Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Child and Family Covariates												
Intercept	-.21	.17	-1.30	.42	-2.32	1.89	-.17	.17	-.99	.50	-2.39	2.05
Female	.26	.05	5.75	.11	-.31	.83	.26	.05	5.77	.11	-.32	.84
Black	-.08	.03	-2.37	.25	-.49	.34	-.07	.03	-1.96	.30	-.50	.36
Hispanic	-.33	.10	-3.33	.19	-1.59	.93	-.33	.09	-3.51	.18	-1.51	.86
Asian	-.02	.07	-.29	.82	-.88	.84	-.03	.07	-.46	.73	-.87	.81
Native Hawaiian or Pacific Islander	.04	.28	.13	.92	-3.48	3.55	.04	.28	.16	.90	-3.46	3.55
Amerian Indian	.15	.16	.97	.51	-1.85	2.16	.13	.15	.84	.55	-1.83	2.09
Not Hispanic	-.22	.04	-5.61	.11	-.72	.28	-.22	.04	-6.21	.10	-.68	.23
Two Races Not Hispanic	.16	.14	1.16	.45	-1.56	1.88	.16	.13	1.26	.43	-1.47	1.79
Receives WIC Benefits	.02	.05	.52	.70	-.56	.61	.03	.05	.57	.67	-.58	.64
Child Age in Months	.06	.00	39.80	.02	.04	.08	.06	.00	41.33	.02	.04	.08
Less Than 35 Hours a Week	-.02	.09	-.19	.88	-1.20	1.16	-.03	.10	-.26	.84	-1.25	1.20
Looking for Work	-.20	.20	-1.01	.50	-2.79	2.38	-.21	.20	-1.06	.48	-2.79	2.36
Not in the Labor Force	.00	.06	.06	.96	-.78	.79	-.01	.06	-.13	.92	-.81	.80
Mother's Age	.00	.00	4.53	.14	-.01	.02	.01	.00	4.59	.14	-.01	.02
Number of Children in House < 18 Years Old	-.08	.02	-4.36	.14	-.33	.16	-.08	.02	-3.88	.16	-.36	.19
High School	.08	.12	.66	.63	-1.44	1.60	.08	.11	.71	.61	-1.35	1.51
Some College	.11	.15	.76	.59	-1.80	2.03	.13	.14	.88	.54	-1.70	1.95
Bachelor	.12	.05	2.21	.27	-.56	.80	.13	.06	2.31	.26	-.58	.83
Graduate	.35	.14	2.57	.24	-1.37	2.07	.35	.14	2.58	.24	-1.38	2.08
Family SES Status	.25	.03	7.45	.09	-.18	.68	.25	.03	7.86	.08	-.15	.65
Suburban	-.23	.03	-6.63	.10	-.68	.21	-.23	.04	-5.58	.11	-.76	.29
Urban	-.12	.13	-.91	.53	-1.82	1.57	-.12	.13	-.90	.53	-1.76	1.52
Teacher and Center Covariates												
Black	-.07	.08	-.81	.57	-1.14	1.01	-.07	.09	-.79	.58	-1.22	1.08
Latino	-.17	.07	-2.39	.25	-1.07	.73	-.17	.07	-2.43	.25	-1.08	.73
Other	-.14	.06	-2.18	.27	-.96	.68	-.14	.06	-2.18	.27	-.94	.67
High School	-.33	.12	-2.85	.22	-1.82	1.15	-.36	.12	-3.04	.20	-1.87	1.15
Vocational	-.36	.16	-2.25	.27	-2.42	1.69	-.39	.13	-2.97	.21	-2.07	1.29
Some College	-.18	.07	-2.66	.23	-1.07	.70	-.21	.07	-2.96	.21	-1.13	.70

Table B3. *Multivariate Regression for Children's Reading Scores for the Provisions for Learning/Teaching Interactions 2-Dimension Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Associate	-.13	.10	-1.35	.41	-1.34	1.08	-.17	.10	-1.63	.35	-1.47	1.14
Bachelor	-.20	.04	-5.47	.12	-.66	.26	-.24	.04	-6.33	.10	-.71	.24
Graduate	-.29	.06	-4.98	.13	-1.02	.44	-.33	.05	-6.68	.10	-.96	.30
Private	.25	.16	1.52	.37	-1.83	2.33	.24	.17	1.46	.38	-1.86	2.34
State/Local	-.09	.04	-2.45	.25	-.53	.36	-.07	.03	-2.57	.24	-.44	.29
Other	-.14	.02	-6.31	.10	-.41	.14	-.12	.02	-7.14	.09	-.35	.10
Program Part Time	-.04	.04	-1.06	.48	-.53	.45	-.03	.04	-.96	.51	-.49	.42
Provisions for Learning Rasch Score							.03	.00	15.26	.04	.01	.06
Teaching and Interactions Rasch Score							.02	.01	1.81	.32	-.15	.20
R2	.40						.40					
f2	.67						.67					

Note. N = 950, and rounded to the nearest 50 per the ECLS-B user agreement. A priori power analysis using GPower version 3.1 (2009) indicated that a sample size of N=83 would be sufficient to detect an effect for the Rasch score, assuming f^2=.10, α = .05, using, 36 predictors, with power = .80. A post hoc sensitivity analysis revealed that given α = .05, power = .80, N = 977 and 37 predictors, there was enough power to detect an effect size of f^2=.01.

Table B4. *Multivariate Regression for Children's Reading Scores for the 2-Dimension Structural/Process Within-Item Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Child and Family Covariates												
Intercept	-.21	.17	-1.30	.42	-2.32	1.89	-.21	.17	-1.26	.43	-2.31	1.89
Female	.26	.05	5.75	.11	-.31	.83	.26	.04	6.25	.10	-.27	.80
Black	-.08	.03	-2.37	.25	-.49	.34	-.07	.03	-2.07	.29	-.49	.36
Hispanic	-.33	.10	-3.33	.19	-1.59	.93	-.33	.09	-3.53	.18	-1.51	.85
Asian	-.02	.07	-.29	.82	-.88	.84	-.03	.06	-.52	.70	-.83	.77
Native Hawaiian or Pacific Islander	.04	.28	.13	.92	-3.48	3.55	.04	.27	.15	.91	-3.45	3.53
Amerian Indian	.15	.16	.97	.51	-1.85	2.16	.12	.15	.76	.59	-1.83	2.06
Not Hispanic	-.22	.04	-5.61	.11	-.72	.28	-.22	.03	-6.78	.09	-.64	.19
Two Races Not Hispanic	.16	.14	1.16	.45	-1.56	1.88	.16	.13	1.29	.42	-1.43	1.75
Receives WIC Benefits	.02	.05	.52	.70	-.56	.61	.03	.05	.56	.68	-.58	.64
Child Age in Months	.06	.00	39.80	.02	.04	.08	.06	.00	42.31	.02	.04	.08
Less Than 35 Hours a Week	-.02	.09	-.19	.88	-1.20	1.16	-.02	.09	-.23	.86	-1.19	1.15
Looking for Work	-.20	.20	-1.01	.50	-2.79	2.38	-.21	.19	-1.08	.48	-2.68	2.26
Not in the Labor Force	.00	.06	.06	.96	-.78	.79	-.01	.06	-.11	.93	-.79	.77
Mother's Age	.00	.00	4.53	.14	-.01	.02	.01	.00	4.68	.13	-.01	.02
Number of Children in House < 18 Years Old	-.08	.02	-4.36	.14	-.33	.16	-.08	.02	-3.92	.16	-.36	.19
High School	.08	.12	.66	.63	-1.44	1.60	.08	.11	.76	.59	-1.33	1.50
Some College	.11	.15	.76	.59	-1.80	2.03	.13	.14	.91	.53	-1.70	1.96
Bachelor	.12	.05	2.21	.27	-.56	.80	.13	.05	2.44	.25	-.56	.82
Graduate	.35	.14	2.57	.24	-1.37	2.07	.36	.13	2.70	.23	-1.32	2.04

Table B4. *Multivariate Regression for Children's Reading Scores for the 2-Dimension Structural/Process Within-Item Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Family SES Status	.25	.03	7.45	.09	-.18	.68	.25	.03	8.18	.08	-.14	.63
Suburban	-.23	.03	-6.63	.10	-.68	.21	-.23	.04	-5.52	.11	-.76	.30
Urban	-.12	.13	-.91	.53	-1.82	1.57	-.12	.12	-.94	.52	-1.70	1.47
Teacher and Center Covariates												
Black	-.07	.08	-.81	.57	-1.14	1.01	-.07	.09	-.84	.56	-1.18	1.03
Latino	-.17	.07	-2.39	.25	-1.07	.73	-.18	.07	-2.47	.25	-1.08	.73
Other	-.14	.06	-2.18	.27	-.96	.68	-.14	.06	-2.23	.27	-.91	.64
High School	-.33	.12	-2.85	.22	-1.82	1.15	-.36	.12	-2.94	.21	-1.91	1.19
Vocational	-.36	.16	-2.25	.27	-2.42	1.69	-.40	.14	-2.82	.22	-2.19	1.39
Some College	-.18	.07	-2.66	.23	-1.07	.70	-.22	.08	-2.84	.22	-1.18	.75
Associate	-.13	.10	-1.35	.41	-1.34	1.08	-.17	.10	-1.65	.35	-1.46	1.13
Bachelor	-.20	.04	-5.47	.12	-.66	.26	-.24	.04	-6.13	.10	-.73	.25
Graduate	-.29	.06	-4.98	.13	-1.02	.44	-.33	.05	-6.46	.10	-.98	.32
Private	.25	.16	1.52	.37	-1.83	2.33	.24	.15	1.57	.36	-1.72	2.20
State/Local	-.09	.04	-2.45	.25	-.53	.36	-.08	.03	-2.72	.22	-.44	.28
Other	-.14	.02	-6.31	.10	-.41	.14	-.13	.02	-6.74	.09	-.37	.11
Program Part Time	-.04	.04	-1.06	.48	-.53	.45	-.03	.03	-.91	.53	-.47	.41
Structural							.03	.00	25.05	.03	.01	.04
Process							.02	.00	39.83	.02	.01	.03
R2	.40						.40					
f2	.67						.67					

Note. N = 950, and rounded to the nearest 50 per the ECLS-B user agreement. A priori power analysis using GPower version 3.1 (2009) indicated that a sample size of N=83 would be sufficient to detect an effect for the Rasch score, assuming f^2=.10, α = .05, using, 36 predictors, with power = .80. A post hoc sensitivity analysis revealed that given α = .05, power = .80, N = 977 and 37 predictors, there was enough power to detect an effect size of f^2=.01.

Table B5. *Multivariate Regression for Children's Math Scores for the 37-Item Unidimensional Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Child and Family Covariates												
Intercept	-.58	.02	-36.38	.02	-.79	-.38	-.58	.02	-27.48	.02	-.85	-.31
Female	.14	.04	3.72	.17	-.33	.61	.14	.04	3.59	.17	-.35	.62
Black	-.01	.04	-.15	.90	-.53	.52	-.01	.04	-.20	.88	-.53	.51
Hispanic	-.10	.00	-22.70	.03	-.16	-.04	-.10	.01	-13.33	.05	-.20	.00
Asian	.09	.11	.81	.57	-1.31	1.49	.09	.11	.82	.56	-1.32	1.50
Native Hawaiian or Pacific Islander	.20	.13	1.58	.36	-1.44	1.85	.20	.13	1.56	.36	-1.44	1.85
Amerian Indian	.40	.23	1.70	.34	-2.57	3.36	.39	.23	1.69	.34	-2.56	3.35
Not Hispanic	-.69	.48	-1.46	.38	-6.73	5.35	-.69	.48	-1.44	.39	-6.76	5.38
Two Races Not Hispanic	.15	.08	1.92	.31	-.84	1.14	.15	.08	1.92	.31	-.84	1.14
Receives WIC Benefits	.03	.07	.48	.72	-.80	.87	.03	.06	.48	.72	-.79	.85
Child Age in Months	.07	.00	58.58	.01	.05	.08	.07	.00	58.25	.01	.05	.08
Less Than 35 Hours a Week	.00	.00	-2.33	.26	-.03	.02	.00	.00	-1.25	.43	-.04	.03

Table B5. *Multivariate Regression for Children's Math Scores for the 37-Item Unidimensional Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Looking for Work	-.25	.05	-5.30	.12	-.83	.34	-.24	.04	-5.58	.11	-.79	.31
Not in the Labor Force	-.01	.05	-.24	.85	-.61	.59	-.01	.05	-.20	.87	-.60	.58
Mother's Age	.01	.01	1.19	.45	-.08	.10	.01	.01	1.19	.45	-.08	.10
Number of Children in House < 18 Years Old	-.07	.05	-1.63	.35	-.65	.50	-.07	.04	-1.65	.35	-.64	.49
High School	.33	.07	4.75	.13	-.56	1.23	.33	.07	4.44	.14	-.61	1.27
Some College	.37	.14	2.58	.24	-1.45	2.19	.36	.15	2.49	.24	-1.49	2.21
Bachelor	.52	.02	31.26	.02	.31	.73	.51	.02	30.26	.02	.30	.73
Graduate	.67	.14	4.75	.13	-1.13	2.48	.67	.14	4.63	.14	-1.17	2.51
Family SES Status	.16	.00	161.67	.00	.15	.18	.16	.00	558.17	.00	.16	.17
Suburban	-.21	.03	-7.40	.09	-.56	.15	-.21	.02	-8.34	.08	-.52	.11
Urban	-.22	.12	-1.86	.31	-1.75	1.30	-.23	.12	-1.85	.32	-1.77	1.32
Teacher and Center Covariates												
Black	-.12	.10	-1.27	.42	-1.35	1.10	-.12	.10	-1.28	.42	-1.34	1.09
Latino	-.08	.04	-2.00	.30	-.57	.42	-.08	.04	-1.75	.33	-.62	.47
Other	-.14	.10	-1.40	.39	-1.40	1.12	-.14	.10	-1.40	.40	-1.42	1.14
High School	-.22	.17	-1.31	.42	-2.35	1.92	-.21	.17	-1.24	.43	-2.40	1.98
Vocational	-.35	.55	-.64	.64	-7.32	6.61	-.34	.56	-.61	.65	-7.45	6.76
Some College	-.17	.20	-.85	.55	-2.73	2.38	-.16	.20	-.80	.57	-2.77	2.44
Associate	-.11	.20	-.54	.69	-2.59	2.38	-.10	.20	-.49	.71	-2.63	2.44
Bachelor	-.14	.16	-.89	.54	-2.19	1.91	-.13	.17	-.80	.57	-2.25	1.99
Graduate	-.18	.11	-1.72	.34	-1.53	1.16	-.17	.11	-1.52	.37	-1.59	1.25
Private	.23	.09	2.54	.24	-.94	1.41	.24	.09	2.57	.24	-.93	1.40
State/Local	-.04	.06	-.62	.65	-.82	.74	-.04	.06	-.67	.63	-.84	.76
Other	-.06	.10	-.58	.67	-1.30	1.18	-.06	.10	-.61	.65	-1.32	1.20
Program Part Time	-.01	.05	-.22	.86	-.64	.62	-.01	.05	-.25	.85	-.65	.63
Rasch Scores							-.01	.00	-2.32	.26	-.07	.05
R2	.36						.36					
f2	.56						.56					

Note. N = 950, and rounded to the nearest 50 per the ECLS-B user agreement. A priori power analysis using GPower version 3.1 (2009) indicated that a sample size of N=82 would be sufficient to detect an effect for the Rasch score, assuming f^2=.10, α = .05, using, 36 predictors, with power = .80. A post hoc sensitivity analysis revealed that given α = .05, power = .80, N = 977 and 36 predictors, there was enough power to detect an effect size of f^2=.01.

Table B6. *Multivariate Regression for Children's Math Scores for the 16-Item Unidimensional Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Child and Family Covariates												
Intercept	-.58	.02	-36.38	.02	-.79	-.38	-.58	.01	-42.36	.02	-.76	-.41
Female	.14	.04	3.72	.17	-.33	.61	.14	.04	3.53	.18	-.36	.64
Black	-.01	.04	-.15	.90	-.53	.52	-.01	.04	-.12	.92	-.58	.57
Hispanic	-.10	.00	-22.70	.03	-.16	-.04	-.10	.00	-41.77	.02	-.13	-.07
Asian	.09	.11	.81	.57	-1.31	1.49	.09	.11	.79	.57	-1.33	1.51
Native Hawaiian or Pacific Islander	.20	.13	1.58	.36	-1.44	1.85	.21	.13	1.63	.35	-1.39	1.80
Amerian Indian	.40	.23	1.70	.34	-2.57	3.36	.39	.23	1.72	.34	-2.53	3.32
Not Hispanic	-.69	.48	-1.46	.38	-6.73	5.35	-.69	.48	-1.45	.38	-6.75	5.37
Two Races Not Hispanic	.15	.08	1.92	.31	-.84	1.14	.15	.08	1.93	.30	-.83	1.13
Receives WIC Benefits	.03	.07	.48	.72	-.80	.87	.03	.07	.47	.72	-.81	.87
Child Age in Months	.07	.00	58.58	.01	.05	.08	.07	.00	60.49	.01	.05	.08
Less Than 35 Hours a Week	.00	.00	-2.33	.26	-.03	.02	-.01	.00	-5.81	.11	-.02	.01
Looking for Work	-.25	.05	-5.30	.12	-.83	.34	-.25	.04	-5.85	.11	-.78	.29
Not in the Labor Force	-.01	.05	-.24	.85	-.61	.59	-.01	.04	-.27	.83	-.58	.55
Mother's Age	.01	.01	1.19	.45	-.08	.10	.01	.01	1.19	.44	-.08	.10
Number of Children in House < 18 Years Old	-.07	.05	-1.63	.35	-.65	.50	-.07	.05	-1.63	.35	-.65	.50
High School	.33	.07	4.75	.13	-.56	1.23	.34	.06	5.29	.12	-.47	1.14
Some College	.37	.14	2.58	.24	-1.45	2.19	.37	.13	2.84	.22	-1.29	2.03
Bachelor	.52	.02	31.26	.02	.31	.73	.52	.03	19.91	.03	.19	.85
Graduate	.67	.14	4.75	.13	-1.13	2.48	.68	.13	5.03	.13	-1.03	2.38
Family SES Status	.16	.00	161.67	.00	.15	.18	.16	.00	217.07	.00	.15	.17
Suburban	-.21	.03	-7.40	.09	-.56	.15	-.21	.03	-7.56	.08	-.55	.14
Urban	-.22	.12	-1.86	.31	-1.75	1.30	-.22	.12	-1.84	.32	-1.77	1.32
Teacher and Center Covariates												
Black	-.12	.10	-1.27	.42	-1.35	1.10	-.12	.10	-1.28	.42	-1.34	1.10
Latino	-.08	.04	-2.00	.30	-.57	.42	-.08	.04	-1.95	.30	-.59	.43
Other	-.14	.10	-1.40	.39	-1.40	1.12	-.14	.10	-1.36	.40	-1.43	1.15
High School	-.22	.17	-1.31	.42	-2.35	1.92	-.22	.18	-1.26	.43	-2.46	2.01
Vocational	-.35	.55	-.64	.64	-7.32	6.61	-.36	.56	-.64	.64	-7.44	6.72
Some College	-.17	.20	-.85	.55	-2.73	2.38	-.17	.21	-.82	.56	-2.87	2.52
Associate	-.11	.20	-.54	.69	-2.59	2.38	-.11	.21	-.52	.69	-2.75	2.54
Bachelor	-.14	.16	-.89	.54	-2.19	1.91	-.15	.17	-.84	.56	-2.36	2.07
Graduate	-.18	.11	-1.72	.34	-1.53	1.16	-.19	.12	-1.51	.37	-1.75	1.38
Private	.23	.09	2.54	.24	-.94	1.41	.23	.09	2.62	.23	-.90	1.37
State/Local	-.04	.06	-.62	.65	-.82	.74	-.04	.06	-.58	.67	-.85	.78
Other	-.06	.10	-.58	.67	-1.30	1.18	-.06	.10	-.55	.68	-1.34	1.23
Program Part Time	-.01	.05	-.22	.86	-.64	.62	-.01	.05	-.20	.88	-.67	.65
Rasch Scores							.00	.01	.25	.85	-.13	.14
R2	.36						.36					
f2	.56						.56					

Note. N = 950, and rounded to the nearest 50 per the ECLS-B user agreement. A priori power analysis using GPower version 3.1 (2009) indicated that a sample size of N=82 would be sufficient to detect an effect for the Rasch score, assuming f^2=.10, α = .05, using, 36 predictors, with power = .80. A post hoc sensitivity analysis revealed that given α = .05, power = .80, N = 977 and 36 predictors, there was enough power to detect an effect size of f^2=.01.

Table B7. *Multivariate Regression for Children's Math Scores for the Provisions for Learning/Teaching Interactions 2-Dimension Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Child and Family Covariates												
Intercept	-.58	.02	-36.38	.02	-.79	-.38	-.53	.03	-19.49	.03	-.87	-.18
Female	.14	.04	3.72	.17	-.33	.61	.14	.04	3.48	.18	-.38	.67
Black	-.01	.04	-.15	.90	-.53	.52	.00	.05	.01	.99	-.69	.69
Hispanic	-.10	.00	-22.70	.03	-.16	-.04	-.10	.00	-54.25	.01	-.13	-.08
Asian	.09	.11	.81	.57	-1.31	1.49	.08	.11	.80	.57	-1.27	1.44
Native Hawaiian or Pacific Islander	.20	.13	1.58	.36	-1.44	1.85	.21	.13	1.61	.35	-1.47	1.90
American Indian	.40	.23	1.70	.34	-2.57	3.36	.41	.21	1.91	.31	-2.31	3.13
Not Hispanic	-.69	.48	-1.46	.38	-6.73	5.35	-.70	.48	-1.47	.38	-6.78	5.37
Two Races Not Hispanic	.15	.08	1.92	.31	-.84	1.14	.15	.09	1.73	.33	-.98	1.28
Receives WIC Benefits	.03	.07	.48	.72	-.80	.87	.03	.07	.37	.77	-.88	.93
Child Age in Months	.07	.00	58.58	.01	.05	.08	.07	.00	63.74	.01	.06	.08
Less Than 35 Hours a Week	.00	.00	-2.33	.26	-.03	.02	-.02	.00	-4.39	.14	-.08	.04
Looking for Work	-.25	.05	-5.30	.12	-.83	.34	-.26	.03	-9.02	.07	-.62	.11
Not in the Labor Force	-.01	.05	-.24	.85	-.61	.59	-.02	.03	-.56	.67	-.44	.40
Mother's Age	.01	.01	1.19	.45	-.08	.10	.01	.01	1.23	.43	-.08	.09
Number of Children in House < 18 Years Old	-.07	.05	-1.63	.35	-.65	.50	-.07	.05	-1.59	.36	-.66	.51
High School	.33	.07	4.75	.13	-.56	1.23	.32	.05	7.17	.09	-.25	.90
Some College	.37	.14	2.58	.24	-1.45	2.19	.36	.11	3.37	.18	-1.00	1.72
Bachelor	.52	.02	31.26	.02	.31	.73	.51	.04	12.21	.05	-.02	1.04
Graduate	.67	.14	4.75	.13	-1.13	2.48	.66	.12	5.42	.12	-.88	2.20
Family SES Status	.16	.00	161.67	.00	.15	.18	.16	.00	39.43	.02	.11	.22
Suburban	-.21	.03	-7.40	.09	-.56	.15	-.20	.02	-9.19	.07	-.49	.08
Urban	-.22	.12	-1.86	.31	-1.75	1.30	-.22	.13	-1.73	.33	-1.86	1.42
Teacher and Center Covariates												
Black	-.12	.10	-1.27	.42	-1.35	1.10	-.11	.09	-1.18	.45	-1.29	1.07
Latino	-.08	.04	-2.00	.30	-.57	.42	-.06	.04	-1.56	.36	-.59	.46
Other	-.14	.10	-1.40	.39	-1.40	1.12	-.14	.10	-1.43	.39	-1.35	1.08
High School	-.22	.17	-1.31	.42	-2.35	1.92	-.20	.19	-1.04	.49	-2.67	2.27
Vocational	-.35	.55	-.64	.64	-7.32	6.61	-.31	.56	-.55	.68	-7.42	6.80
Some College	-.17	.20	-.85	.55	-2.73	2.38	-.14	.25	-.57	.67	-3.27	2.99
Associate	-.11	.20	-.54	.69	-2.59	2.38	-.09	.25	-.34	.79	-3.27	3.09
Bachelor	-.14	.16	-.89	.54	-2.19	1.91	-.12	.21	-.56	.67	-2.77	2.53
Graduate	-.18	.11	-1.72	.34	-1.53	1.16	-.17	.16	-1.04	.49	-2.20	1.86
Private	.23	.09	2.54	.24	-.94	1.41	.22	.08	2.86	.21	-.74	1.18
State/Local	-.04	.06	-.62	.65	-.82	.74	-.03	.06	-.41	.75	-.82	.77
Other	-.06	.10	-.58	.67	-1.30	1.18	-.05	.10	-.48	.72	-1.38	1.28
Program Part Time	-.01	.05	-.22	.86	-.64	.62	-.02	.06	-.33	.80	-.84	.79

Table B7. *Multivariate Regression for Children's Math Scores for the Provisions for Learning/Teaching Interactions 2-Dimension Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Provisions for Learning Rasch Scores							-.05	.05	-.91	.53	-.71	.61
Teaching and Interactions Rasch Scores							.08	.00	19.86	.03	.03	.14
R2	.36						.36					
f2	.56						.56					

Note. N = 950, and rounded to the nearest 50 per the ECLS-B user agreement. A priori power analysis using GPower version 3.1 (2009) indicated that a sample size of N=82 would be sufficient to detect an effect for the Rasch score, assuming f^2=.10, α = .05, using, 36 predictors, with power = .80. A post hoc sensitivity analysis revealed that given α = .05, power = .80, N = 977 and 36 predictors, there was enough power to detect an effect size of f^2=.01.

Table B8. *Multivariate Regression for Children's Math Scores for the Structural/Process 2-Dimension Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Child and Family Covariates												
Intercept	-.58	.02	-36.38	.02	-.79	-.38	-.59	.01	-88.45	.01	-.68	-.51
Female	.14	.04	3.72	.17	-.33	.61	.14	.04	3.74	.17	-.34	.63
Black	-.01	.04	-.15	.90	-.53	.52	-.01	.05	-.15	.91	-.64	.63
Hispanic	-.10	.00	-22.70	.03	-.16	-.04	-.10	.00	-103.91	.01	-.11	-.09
Asian	.09	.11	.81	.57	-1.31	1.49	.08	.11	.74	.59	-1.36	1.53
Native Hawaiian or Pacific Islander	.20	.13	1.58	.36	-1.44	1.85	.21	.13	1.68	.34	-1.39	1.81
Amerian Indian	.40	.23	1.70	.34	-2.57	3.36	.39	.21	1.85	.32	-2.29	3.07
Not Hispanic	-.69	.48	-1.46	.38	-6.73	5.35	-.70	.47	-1.49	.38	-6.63	5.24
Two Races Not Hispanic	.15	.08	1.92	.31	-.84	1.14	.15	.09	1.77	.33	-.93	1.23
Receives WIC Benefits	.03	.07	.48	.72	-.80	.87	.03	.07	.37	.78	-.87	.92
Child Age in Months	.07	.00	58.58	.01	.05	.08	.07	.00	71.25	.01	.06	.08
Less Than 35 Hours a Week	.00	.00	-2.33	.26	-.03	.02	-.01	.00	-45.50	.01	-.02	-.01
Looking for Work	-.25	.05	-5.30	.12	-.83	.34	-.25	.04	-5.76	.11	-.80	.30
Not in the Labor Force	-.01	.05	-.24	.85	-.61	.59	-.02	.04	-.42	.75	-.47	.44
Mother's Age	.01	.01	1.19	.45	-.08	.10	.01	.01	1.23	.44	-.08	.09
Number of Children in House < 18 Years Old	-.07	.05	-1.63	.35	-.65	.50	-.07	.05	-1.58	.36	-.66	.51
High School	.33	.07	4.75	.13	-.56	1.23	.34	.05	6.85	.09	-.29	.96
Some College	.37	.14	2.58	.24	-1.45	2.19	.37	.12	3.15	.20	-1.13	1.87
Bachelor	.52	.02	31.26	.02	.31	.73	.52	.03	15.97	.04	.11	.93
Graduate	.67	.14	4.75	.13	-1.13	2.48	.68	.13	5.33	.12	-.93	2.28
Family SES Status	.16	.00	161.67	.00	.15	.18	.16	.00	227.22	.00	.15	.17
Suburban	-.21	.03	-7.40	.09	-.56	.15	-.20	.02	-10.69	.06	-.45	.04
Urban	-.22	.12	-1.86	.31	-1.75	1.30	-.23	.12	-1.87	.31	-1.79	1.33
Teacher and Center Covariates												

Table B8. *Multivariate Regression for Children's Math Scores for the Structural/Process 2-Dimension Specification of the Measure*

Parameters	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%	Coef.	Std. Err.	t	Sig.(*p*)	Upper 95%	Lower 95%
Black	-.12	.10	-1.27	.42	-1.35	1.10	-.11	.10	-1.16	.45	-1.34	1.11
Latino	-.08	.04	-2.00	.30	-.57	.42	-.07	.04	-2.05	.29	-.53	.38
Other	-.14	.10	-1.40	.39	-1.40	1.12	-.14	.10	-1.38	.40	-1.41	1.13
High School	-.22	.17	-1.31	.42	-2.35	1.92	-.21	.19	-1.12	.46	-2.56	2.15
Vocational	-.35	.55	-.64	.64	-7.32	6.61	-.33	.55	-.60	.66	-7.35	6.68
Some College	-.17	.20	-.85	.55	-2.73	2.38	-.16	.22	-.70	.61	-2.99	2.68
Associate	-.11	.20	-.54	.69	-2.59	2.38	-.10	.23	-.44	.74	-2.96	2.76
Bachelor	-.14	.16	-.89	.54	-2.19	1.91	-.13	.19	-.68	.62	-2.50	2.25
Graduate	-.18	.11	-1.72	.34	-1.53	1.16	-.17	.13	-1.31	.42	-1.87	1.52
Private	.23	.09	2.54	.24	-.94	1.41	.22	.09	2.43	.25	-.92	1.35
State/Local	-.04	.06	-.62	.65	-.82	.74	-.04	.06	-.58	.66	-.83	.76
Other	-.06	.10	-.58	.67	-1.30	1.18	-.06	.10	-.58	.67	-1.37	1.25
Program Part Time	-.01	.05	-.22	.86	-.64	.62	-.02	.05	-.29	.82	-.69	.66
Structural							-.02	.02	-1.01	.50	-.24	.21
Process							.03	.02	1.57	.36	-.20	.26
R2	.36						.36					
f2	.56						.56					

Note. N = 950, and rounded to the nearest 50 per the ECLS-B user agreement. A priori power analysis using GPower version 3.1 (2009) indicated that a sample size of N=82 would be sufficient to detect an effect for the Rasch score, assuming f^2=.10, α = .05, using, 36 predictors, with power = .80. A post hoc sensitivity analysis revealed that given α = .05, power = .80, N = 977 and 36 predictors, there was enough power to detect an effect size of f^2=.01.