DISSERTATION

THE VALIDITY AND PREDICTIVE UTILITY OF THE CLASSROOM ASSESSMENT SCORING SYSTEM PRE-K IN HEAD START CLASSROOMS

Submitted by

Caitlyn Grubb

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2025

Doctoral Committee:

Advisor: Kim Henry

Brad Connor
Emily Merz
Lisa Daunhauer

ABSTRACT

THE VALIDITY AND PREDICTIVE UTILITY OF THE CLASSROOM ASSESSMENT SCORING SYSTEM PRE-K IN HEAD START CLASSROOMS

In this dissertation, I critically examine the measurement structure and predictive validity of the Classroom Assessment Scoring System Pre-K (CLASS Pre-K) within Head Start settings, with implications for research, practice, and policy. Drawing on a comprehensive review of empirical studies, the analysis reveals that both the traditional three-factor model and the bifactor model of CLASS Pre-K exhibit poor model fit, high inter-factor correlations, and inconsistent item loadings, raising concerns about their structural validity in diverse classrooms. Additionally, CLASS Pre-K scores show inconsistent and modest associations with child outcomes, particularly in socio-emotional development. These limitations reflect deeper conceptual shortcomings in the framework, including its narrow scope and limited cultural responsiveness. In high-need environments like Head Start, quality teaching involves trauma-informed care, culturally relevant pedagogy, and family engagement—dimensions largely absent from CLASS metrics. In this dissertation, I argue that reliance on CLASS Pre-K in high-stakes accountability systems such as the Designation Renewal System may inadvertently penalize programs serving marginalized communities. To improve equity and accuracy in classroom quality assessment, this work recommends revising the CLASS framework to include culturally grounded and contextually relevant indicators. It calls for participatory, mixed-methods research and subgroup-specific validation studies to develop more responsive tools that align with the strengths and realities of diverse early learning setting

TABLE OF CONTENTS

THE VALIDITY AND PREDICTIVE UTILITY OF THE CLASSROOM ASSESSMENT SCORING SYSTEM PRE-K IN HEAD START CLASSROOMS

The Critical Importance of Quality Experiences in Early Childhood Education

The classroom experience is a cornerstone of children's cognitive, socio-emotional, and academic development during their pre-kindergarten years (Peisner-Feinberg et al., 2001; Wang et al., 2020). Pre-K programs are early childhood education initiatives for children ages 3 to 5 that aim to prepare them for kindergarten by fostering cognitive, social-emotional, and physical development, supporting foundational skills in language, literacy, numeracy, and self-regulation, and are offered through various public and private settings with funding from federal, state, or local sources (Mitchell, 2001). High-quality pre-K classrooms are characterized by strong teacher-child interactions, effective routines, and supportive relationships, all of which contribute to fostering a positive learning environment (Tatalović Vorkapić & LoCasale-Crouch, 2021). Early education provides a critical foundation for academic success, with evidence highlighting the impact of classroom quality, encompassing structural features, teacher qualifications, and process quality, on children's early learning experiences (Burchinal et al., 2010; Ulferts et al., 2019). Beyond resources and aesthetics, classroom quality reflects a dynamic interplay of teacher expertise, interactive practices, and child experiences that create conducive settings for growth and learning (Pianta et al., 2016). Research consistently underscores that nurturing, stimulating, and well-structured classrooms catalyze significant cognitive, social, and emotional development in young learners (Abreu-Lima et al., 2013; Anders et al., 2012; Auger et al., 2014; Burchinal et al., 2010; Campbell et al., 2002; Mashburn et al., 2008; Rimm-Kaufman et al., 2005; Wang et al., 2020; Zheng et al., 2023).

Classroom quality is a critical factor in shaping early learning experiences and is widely recognized as a key determinant of children's academic and developmental trajectories, with research consistently demonstrating that early childhood education fosters essential cognitive, emotional, and social skills that influence long-term academic and developmental outcomes (Anders et al., 2012; Burchinal et al., 2010; Nores & Barnett, 2010; Peisner-Feinberg et al., 2001). Pre-K programs not only provide an early start to learning but also equip children with tools and strategies to enhance their performance and development (Burchinal et al., 2010; Peisner-Feinberg et al., 2001). Empirical evidence highlights the lasting impact of high-quality early learning environments on children's academic and social trajectories (Magnuson et al., 2007; Ulferts et al., 2019). In that, children who attend high-quality early childhood education programs demonstrate stronger school readiness, higher academic achievement, and more positive behavioral outcomes than those in lower-quality programs. (Burchinal et al., 2010; Mashburn et al., 2008; Nores & Barnett, 2010). Moreover, the benefits extend beyond the early years, with longitudinal research linking high-quality pre-K to higher educational attainment, reduced behavioral challenges, and improved socioeconomic outcomes in adulthood (Camilli et al., 2010; Magnuson et al., 2007; Nores & Barnett, 2010). These findings emphasize the necessity of maintaining program quality to support lifelong learning and child development.

High-quality Classrooms are Important for Children from Disadvantaged Families

While a high-quality classroom is important for all children, it may be especially important for children from low-income families (Gormley et al., 2005; Weiland & Yoshikawa, 2013; Berkowitz et al., 2017; Moen et al., 2019). Children from disadvantaged backgrounds often face socioeconomic challenges that can hinder their educational advancement, including limited access to high-quality early learning environments, exposure to adverse conditions, and disparities in school climate (Berkowitz et al., 2017; Moen et al., 2019). While research demonstrates that high-quality pre-K programs can help mitigate some of these challenges and improve cognitive and socio-emotional outcomes, their effectiveness varies depending on program quality and implementation (Gormley et al., 2005; Weiland & Yoshikawa, 2013). For many children from disadvantaged backgrounds, programs like Head Start often provide critical early exposure to structured education and academic support, addressing gaps in early learning opportunities and fostering developmental growth (Gormley et al., 2005; Weiland & Yoshikawa, 2013; Moen et al., 2019; Zhai et al., 2014). It is within these classrooms that high-quality instruction and nurturing relationships have the ability to shape children's academic, social, and emotional development (Burchinal et al., 2010; Mashburn et al., 2008; Gormley et al., 2005; Weiland & Yoshikawa, 2013). For this reason, it is crucial to consider the diverse characteristics of children and teachers in the classroom when evaluating measures of quality, as these factors influence the effectiveness and relevance of quality assessments (Pianta et al., 2020; Gormley et al., 2005; Weiland & Yoshikawa, 2013).

Research consistently demonstrates that access to high-quality classrooms is particularly transformative for children from low-income families, who are disproportionately impacted by educational disparities, such as larger class sizes or outdated materials (Gormley et al., 2005; Weiland & Yoshikawa, 2013; Berkowitz et al., 2017; Moen et al., 2019). These preschool years not only shape immediate learning outcomes but also have a significant influence on long-term academic achievement and life trajectories (Berkowitz et al., 2017; Gormley et al., 2005). Supportive and stimulating learning environments are found to mitigate the adverse effects of poverty on educational success (Weiland & Yoshikawa, 2013; Moen et al., 2019). Greater emphasis must be placed on understanding and addressing the specific needs of children from low-income families through research, intentional improvements in classroom quality, and tailored educational practices (Gormley et al., 2005; Weiland & Yoshikawa, 2013). By doing so, we can empower these children, disrupt the cycle of disadvantages, and foster a future where every child can succeed.

HEAD START BRINGS HIGH-QUALITY EDUCATION TO CHILDREN FROM DISADVANTAGED FAMILIES

Head Start began in 1965 as part of President Lyndon B. Johnson's *War on Poverty*, created to promote school readiness for young children from low-income families in the United States (Administration for Children and Families, 2024a). Head Start provides comprehensive services—including education, health care, nutrition, and parental involvement—to support the development and well-being of children from disadvantaged families (Administration for Children and Families, 2024a). This broad approach was designed to help break the cycle of poverty by preparing children for school and addressing their holistic needs from an early age (Administration for Children and Families, 2024a).

From its inception, Head Start demonstrated immediate benefits in children's learning and development, particularly in language, literacy, and math (Garces et al., 2002). However, as early as 1966, researchers observed that these initial academic gains often faded by early elementary school, with test scores between Head Start participants and non-participants becoming similar over time (Camilli et al., 2010). Later evaluations confirmed that without continued high-quality education, many of the program's early advantages diminished by third grade (Weiland & Yoshikawa, 2013). These findings fueled debate about Head Start's long-term value, even as the program expanded to serve nearly one million children annually in subsequent decades (Administration for Children and Families, 2024a).

Despite concerns about academic fade-out, evidence supports Head Start's lasting impact (Gormley et al., 2005), including higher high school graduation rates, improved health outcomes, and increased lifetime earnings (Magnuson & Waldfogel, 2005). Longitudinal studies tracking participants into adolescence and adulthood reveal that Head Start alumni are significantly more likely to graduate from high school and pursue higher education compared to their peers who did not attend (Garces et al., 2002). Beyond academic outcomes, Head Start has also been linked to better health and social-emotional development, contributing to greater economic mobility in adulthood (Fiscella & Kitzman, 2009). These long-term benefits are particularly pronounced among Black children, suggesting the program's potential for positive effects in vulnerable populations (Joshi et al., 2016). Findings suggest that Head Start's impact extends well beyond early academics, providing lasting benefits that support children's lifelong success (Ladd & Goertz, 2015).

Research suggests that sustaining Head Start's early gains requires strong continuity and support as children transition into elementary school (Sabol & Pianta, 2014). One reason early academic advantages may fade is that many Head Start children attend under-resourced K–12 schools, which often struggle to build on the skills gained in pre-K (Valentino, 2018). To counter this, experts advocate aligning pre-K curricula with kindergarten and the early elementary grades so that the progress made in Head Start is reinforced and advanced rather than repeated or lost (Weiland & Yoshikawa, 2013). Strengthening early school instruction and providing continued resources ensures that former Head Start children maintain their academic momentum, closing the achievement gap with their more advantaged peers as they advance through school (Magnuson & Waldfogel, 2005). By implementing high-quality early learning programs beyond pre-K, policymakers and educators can amplify the long-term benefits of Head Start (Temple et al., 2022).

Dual Language Learners

Head Start serves not only low-income families but also an increasingly diverse population, including a growing proportion of Dual Language Learners (DLLs) whose academic and linguistic needs call for targeted, evidence-based instructional strategies (Office of Head Start, 2024b; Valentino, 2018; Sabol, 2014). In 2023, DLLs made up 32.5% of Head Start enrollment—a 12.5% increase since 2019—highlighting the pressing need for classrooms that support both linguistic and cultural development (Office of Head Start, 2024b). The program's staffing capacity, with over 30% of personnel fluent in a language other than English, offers an opportunity to deliver culturally and linguistically responsive instruction (Office of Head Start, 2024b). However, despite this potential, DLLs continue to face unequal access to high-quality, research-based learning environments (Valentino, 2018; Sabol, 2014).

Research emphasizes that language acquisition and academic success for DLLs depend on instructional practices that are not only developmentally appropriate but also tailored to their dual-language contexts (Valentino, 2018; Sabol, 2014). Head Start prioritizes bilingual development as a means of supporting broader cognitive and academic growth, recognizing that strong language skills are foundational for learning across domains (U.S. Department of Health and Human Services, 2022). Nevertheless, many DLLs still lack adequate support in both English and their home language, which can lead to long-term academic disparities (McFarland et al., 2018). By fourth grade, DLLs score an average of 37 points lower in reading compared to monolingual peers, a gap often rooted in early language exposure and access to learning resources (McFarland et al., 2018). Compounding this challenge, children in poverty frequently have limited access to books, educational toys, and other essential tools for vocabulary development, contributing to lower language skills at kindergarten entry (Bulotsky-Shearer et al., 2012; Hindman et al., 2010; Lee & Burkam, 2002).

High-quality early education programs like Head Start are vital for closing these language development gaps, especially for DLLs in low-income households (U.S. Department of Health and Human Services, 2022). Studies show that instructional quality plays a critical role in vocabulary growth in both Spanish and English, particularly for DLLs entering with lower language skills (Hindman & Wasik, 2015). Importantly, the quality of teacher-child interactions has been found to exert a stronger influence on language outcomes than the frequency of vocabulary instruction alone, underscoring the importance of responsive and intentional teaching practices (Hindman & Wasik, 2013). Ensuring that Head Start classrooms provide linguistically responsive, high-quality instruction is therefore essential for supporting bilingual development, reducing early disparities, and promoting long-term academic success (Hindman & Wasik, 2013; Hindman & Wasik, 2015; U.S. Department of Health and Human Services, 2022).

Head Start Designation Renewal System

To ensure the delivery of high-quality services for all children, the Improving Head Start for School Readiness Act of 2007 established rigorous standards for program operations (Department of Health and Human Services, 2020). This legislation led to the implementation of the Designation Renewal System (DRS), which introduced performance-based criteria for continued funding, including meeting school readiness goals, maintaining licensing, demonstrating financial stability, and achieving benchmark scores on the Classroom Assessment Scoring System Pre-K (CLASS Pre-K; Department of Health and Human Services, 2020; Office of Head Start, 2024a). By enforcing these standards, Head Start aims to provide equitable, high-quality early education for children from disadvantaged backgrounds, fulfilling its mission to reduce disparities in educational outcomes (Department of Health and Human Services, 2020).

CLASS Pre-K, developed by Pianta et al. (2008b), is the primary observational measure used in the DRS to assess classroom quality. This tool evaluates teacher-child interactions across three key domains: Emotional Support, Classroom Organization, and Instructional Support.

Scores for each domain range from 1 to 7, with higher scores indicating stronger interactions that promote child development (Pianta et al., 2008b). The DRS requires Head Start programs to meet minimum CLASS Pre-K thresholds to maintain funding: a score of at least 5.0 in Emotional Support, 5.0 in Classroom Organization, and 2.5 in Instructional Support (Department of Health and Human Services, 2020). Programs failing to meet these thresholds must reapply for funding, potentially risking the loss of their grant (Office of Head Start, 2024a).

The DRS assumes that classrooms within the same Head Start center maintain similar levels of quality, which serves as the basis for its accountability framework (Department of Health and Human Services, 2020). Based on this assumption, random classrooms are selected annually for CLASS Pre-K assessments, and the center's average score determines compliance (Department of Health and Human Services, 2020). Large-scale early childhood education studies, such as the National Center for Early Development and Learning (NCEDL) Multi-State Study of Pre-Kindergarten, the Head Start Family and Child Experiences Survey (FACES), and the State-Wide Early Education Program Study (SWEEP), continue to aggregate CLASS Pre-K scores at the center level for assessment and policy decisions (Bustamante & Hindman, 2019). These studies often include only one or a small number of classrooms per center, which may obscure important within-center variations in quality (Sabol et al., 2020). Although research supports the role of center-level factors in shaping classroom quality, studies also indicate substantial within-center variation (Bloom & Weiland, 2015; Sabol et al., 2020). This raises concerns about the accuracy of center-wide assessments (Sabol et al., 2020).

Beyond its role in program accountability, the DRS has generated an extensive dataset containing CLASS Pre-K quality measures and child academic achievement data (Department of Health and Human Services, 2020; Office of Head Start, 2023). While this dataset is valuable for evaluating program effectiveness, evidence of within-center variability in classroom quality suggests the need for more nuanced assessment methods (Bloom & Weiland, 2015; Sabol et al., 2020). Nevertheless, the breadth of data collected through the DRS offers researchers an opportunity to explore key relationships between classroom quality, child outcomes, and program characteristics, contributing to the continuous improvement of Head Start programs (Bustamante & Hindman, 2019; Sabol et al., 2020; Bloom & Weiland, 2015).

Emotional Support

The domain of Emotional Support evaluates the quality of teacher-child interactions, emphasizing emotional tone, sensitivity, and responsiveness, which collectively fosters a supportive and nurturing classroom environment (Pianta et al., 2008b). The theoretical foundation of Emotional Support is grounded in attachment theory (Ainsworth et al., 1978; Bowlby, 1982; Pianta, 1999), which emphasizes the significance of emotional relationships between children and caregivers, extended in this context to child-teacher interactions. The theoretical foundation also includes self-determination theory, which underscores the importance of emotional support as a mechanism for motivating children, focusing on autonomy, competence, and relatedness as critical needs (Ryan & Deci, 2000; Skinner & Belmont, 1993). The Emotional Support domain in the CLASS framework comprises four dimensions: *Negative Climate, Positive Climate, Teacher Sensitivity, and Regard for Child Perspectives* (Pianta et al., 2008b).

Negative Climate and *Positive Climate,* key dimensions of CLASS Pre-K, assess the emotional tone and quality of interactions between teachers, children, and peers, capturing both supportive and adverse elements of classroom environments (Hamre & Pianta, 2010). *Negative Climate* captures indicators such as anger, hostility, or conflict, whereas *Positive Climate* reflects warmth, mutual respect, and emotional connection in the classroom (Hamre & Pianta, 2010; Pianta et al., 2008b). These dimensions are inversely related: teachers receive higher scores for Positive Climate when they foster supportive interactions, whereas Negative Climate is based on the presence of less supportive or more adverse interactions (Hamre & Pianta, 2010; Pianta et al., 2008b). Additionally, these dimensions encompass peer dynamics, including aggression, rejection, or victimization, which are critical markers of negative classroom interactions (Hamre & Pianta, 2010; Pianta et al., 2008b).

Classroom climate is critical to children's development, as research shows that early experiences of peer rejection and victimization are linked to increased risks of academic struggles, school avoidance, and social difficulties, with long-term implications for school adjustment and psychological well-being (Buhs & Ladd, 2001; Buhs et al., 2006; Cassidy & Asher, 1992; Ladd et al., 1997; Ladd & Troop-Gordon, 2003). In contrast, classrooms characterized by positive and emotionally supportive relationships between teachers and peers foster improved social-emotional and academic outcomes by enhancing children' engagement, motivation, and school adjustment (Birch & Ladd, 1996; Crosnoe et al., 2004; Hamre & Pianta, 2005). A supportive classroom climate, grounded in these positive interactions, serves as a critical component of classroom quality, influencing children's long-term developmental and educational trajectories (Hamre & Pianta, 2010; Birch & Ladd, 1996).

Teacher Sensitivity measures the degree to which teachers are attuned to the individual needs of children, fostering a classroom environment where children feel safe, supported, and free to explore and learn (Pianta et al., 2008b). Sensitive teachers exhibit consistent and responsive interactions that contribute to a nurturing atmosphere (Pianta et al., 2008b). Classrooms with high levels of teacher sensitivity are often characterized by greater child engagement, increased self-reliance, and lower levels of externalizing behaviors at home (Rimm-Kaufman et al., 2002). Furthermore, classrooms characterized by warm, responsive teacher-child interactions foster the development of advanced language skills, reinforcing the critical role of teacher sensitivity in supporting both academic and social growth (Connor et al., 2005).

Regard for Child Perspectives, often conceptualized as autonomy support, emphasizes the extent to which teachers follow children' interests and encourage their ideas and suggestions in the classroom (Pianta et al., 2008b; Skinner & Belmont, 1993). This dimension also reflects the balance between teacher-led activities and those guided by child preferences, underscoring a child-centered approach to classroom dynamics (Hamre & Pianta, 2010). Classrooms that score high in this domain foster positive feelings about school among children and enhance their engagement and motivation to learn (Gutman & Sulzby, 2000; Pianta et al., 2002; Valeski & Stipek, 2001). Child-centered environments, which prioritize the interests and psychological needs of children, are associated with more favorable academic and social outcomes, reinforcing the critical role of autonomy support in early childhood education (Gutman & Sulzby, 2000; Hamre & Pianta, 2010; Pianta et al., 2008b; Pianta et al., 2002; Skinner & Belmont, 1993; Valeski & Stipek, 2001).

Importance of High-Quality Emotional Support

A high-quality, emotionally supportive classroom climate is essential for fostering child success, particularly for at-risk pre-K children (Hamre & Pianta, 2005; Moen et al., 2019). Emotionally supportive classrooms and strong teacher-child relationships are critical in shaping positive developmental and academic outcomes, particularly for children from disadvantaged backgrounds (Hamre & Pianta, 2005; Moen et al., 2019). These interactions foster cognitive, emotional, and social growth (Curby et al., 2009a; Mashburn et al., 2008) while protecting against developmental challenges, behavioral difficulties, and the adverse effects of socioeconomic disadvantage (Berkowitz et al., 2017; Burchinal et al., 2002; Qi et al., 2020).

Instructional Support

While Emotional Support focuses on teacher-child relationships, the Instructional Support domain of CLASS Pre-K emphasizes the way teachers present concepts in an organized, interconnected manner that builds upon prior knowledge (Hamre & Pianta, 2010; Pianta et al., 2008b). Instructional Support draws from theories emphasizing the role of scaffolding provided by adults during learning opportunities, enabling children to extend beyond their current abilities (Davis & Miyake, 2004; Skibbe et al., 2004). The CLASS framework operationalizes Instructional Support through three dimensions: *Concept Development, Quality of Feedback,* and *Language Modeling*, each of which is integral to fostering children's engagement and higher-order thinking (Hamre & Pianta, 2010; Pianta et al., 2008b).

Concept Development emphasizes encouraging children to engage in higher order thinking rather than merely memorizing facts (Pianta et al., 2008b). This approach promotes deeper learning, as demonstrated by research showing that children in classrooms focusing on conceptual understanding achieve greater gains in academic performance (Taylor et al., 2003; Wharton-McDonald et al., 1998). A critical aspect of this dimension is the teacher's ability to present information in ways that make learning meaningful and relevant, fostering connections between learning material and children' experiences (Hamre & Pianta, 2010; Davis & Miyake, 2004).

Quality of Feedback focuses on how teachers provide feedback that fosters growth and understanding rather than merely evaluating correctness (Pianta et al., 2008b). This dimension emphasizes process-oriented feedback that guides children toward understanding how to arrive at the correct answers, making learning more meaningful and engaging (Pianta et al., 2008b). High-quality feedback, characterized by responsive teacher-child interactions, often involves questioning strategies that encourage deeper comprehension, a practice aligned with scaffolded instruction (Pianta et al., 2008b). Moreover, process-oriented feedback in pre-K and kindergarten settings has been shown to reduce disparities in school readiness and help close the achievement gap for children from disadvantaged families entering elementary school (Bustamante & Hindman, 2019; Hamre & Pianta, 2005).

Language Modeling refers to teachers employing language-stimulating techniques, such as open-ended questions, repetition, and the use of advanced vocabulary, to support children in expanding their language skills (Pianta et al., 2008b). High-quality and frequent language use by caregivers significantly influences young children’s language development, with convincing evidence linking caregiver speech to vocabulary growth and early cognitive skills, particularly through direct interactions and child-directed speech (Hoff, 2003; Weisleder & Fernald, 2013). *Language Modeling* can impact development by using responsive and intentional language modeling techniques, such as providing explicit instruction and engaging in rich verbal interactions with children (Cabell et al., 2011; Pinto et al., 2013).

Importance of High-Quality Instructional Support

High levels of Instructional Support play a critical role in moderating the relationship between difficult temperament and academic outcomes, significantly reducing its negative impact and helping children with more challenging temperaments achieve similar academic success as their peers (Curby et al., 2011). Research consistently finds that classrooms serving higher concentrations of minority and low-income children receive lower scores on Instructional Support, highlighting systemic inequities in early educational opportunities (Bassok & Galdo, 2016; Pianta et al., 2005; Valentino, 2018). These findings highlight the need for targeted interventions to improve Instructional Support in underserved communities, as research consistently demonstrates that low-income and minority children are more likely to experience lower-quality instructional interactions, reinforcing systemic inequities in educational opportunities (Bassok & Galdo, 2016; Pianta et al., 2005; Valentino, 2018).

Classroom Organization

The final domain, Classroom Organization, assesses teachers' ability to effectively manage children' time, attention, and behavior, ensuring smooth transitions and maximizing instructional time (Pianta et al., 2008b). Effective classroom management strategies contribute to achieving instructional goals by fostering structured, well-organized learning environments (Emmer & Stough, 2001). This factor emphasizes the development of children's self-regulatory skills, including memory, attention, planning, and inhibitory control, which are foundational to both academic and social success (Blair, 2002; Paris & Paris, 2001; Raver, 2004). Pre-K classrooms have been identified as particularly influential in this domain due to the rapid brain development and plasticity that occur during early childhood, supporting significant growth in these areas (Blair, 2002). The Classroom Organization domain is composed of three interrelated dimensions: *Behavior Management, Productivity, and Instructional Learning Formats*(Pianta et al., 2008b).

Behavior Management refers to teachers' ability to establish and maintain a classroom environment that encourages positive behaviors while efficiently preventing or addressing negative behaviors (Pianta et al., 2008b). Effective *Behavior Management* involves setting clear expectations, proactively addressing issues, using positive reinforcement, and redirecting misbehavior with minimal disruption (Emmer & Stough, 2001; Pianta et al., 2008b). Teachers who consistently implement these strategies are more likely to foster child engagement, as well-managed classrooms minimize disruptions and enhance focus on learning tasks (Emmer &

Stough, 2001; Tingstrom et al., 2006). In early childhood education, effective behavior management fosters a structured and supportive classroom environment that helps children develop self-regulatory skills (Hamre & Pianta, 2010).

Productivity refers to the extent to which classroom time is efficiently used for learning rather than management activities like taking attendance (Pianta et al., 2008b). In productive classrooms, teachers skillfully manage transitions and set clear expectations, ensuring minimal loss of instructional time (Hamre & Pianta, 2010; Pianta et al., 2008b). Well-organized classrooms foster child engagement in meaningful learning activities, which enhances academic achievement by maximizing instructional time and reducing disruptions (Cameron et al., 2008). Research indicates that when children spend more time actively engaged in learning, they are more likely to experience better academic outcomes, particularly in literacy and other foundational skills (Cameron et al., 2008). The focus on *Productivity* highlights the critical relationship between effective classroom management and maximizing time-on-task, which is essential for fostering child success (Cameron et al., 2008; Emmer & Stough, 2001; Hamre & Pianta, 2010; Pianta et al., 2008b).

Instructional Learning Formats refers to children being effectively and actively engaged in the learning activities provided in the classroom (Pianta et al., 2008b). High-quality *Instructional Learning Formats* provide children with a structured environment that encourages active participation in educational content while minimizing distractions (Hamre & Pianta, 2010). Classroom environments that actively engage children through *Instructional Learning Formats* promote deeper understanding and academic achievement, yet these practices remain underutilized in early elementary settings (Pianta et al., 2008b; Stites & Brown, 2021). Similar findings in pre-K settings indicate that teachers often spend less time using active learning methods and more time relying on traditional, less engaging approaches (Stites & Brown, 2021). Research emphasizes the need for strategies to increase the use of high-quality instructional formats that focus on active learning, which can foster better engagement and achievement for young children (Cameron et al., 2008; Hamre et al., 2014).

Importance of High-Quality Classroom Organization

Classroom Organization is a critical factor in promoting positive child outcomes, especially for children in vulnerable populations, as well-managed environments support engagement, emotional regulation, and academic success through consistent routines and clear expectations (Gaskins et al., 2012; Rimm-Kaufman et al., 2009). High-quality teacher-child interactions, particularly in Early Head Start settings, further enhance these effects (Choi et al., 2019). Research links effective classroom management to improved behavioral and cognitive self-control, greater engagement, and reduced off-task behavior (Rimm-Kaufman et al., 2009; Cameron et al., 2008), while structured learning environments also support early literacy development by minimizing distractions and enhancing instruction (Cameron et al., 2008; Hamre & Pianta, 2010). These benefits are especially important for struggling children, who are more likely to stay engaged and avoid falling behind in well-organized classrooms (Ferretti, 2014; Gaskins et al., 2012). Despite this evidence, further research is needed to better understand how Classroom Organization affects children from low-income families and to identify strategies for maximizing its impact (Valentino, 2018).

In addition to the CLASS Pre-K observational assessments of classroom quality, Head Start evaluation initiatives focus on assessing behavioral and academic development, as outlined in the Head Start Early Learning Outcomes Framework (ELOF; Office of Head Start, 2015). The ELOF identifies five essential outcome domains critical for school readiness and long-term success: approaches to learning, social and emotional development, language and literacy, cognition, and perceptual, motor, and physical development (Office of Head Start, 2015). These outcome domains are interconnected and mutually influential, forming a comprehensive foundation for understanding children's developmental progress (Office of Head Start, 2015). Within each outcome domain, the ELOF provides specific indicators that describe observable skills, behaviors, and concepts children are expected to develop by the end of their Head Start experience (Office of Head Start, 2015). The ELOF offers valuable developmental insights for parents and educators, serving as a guide for individualizing teaching practices and program planning (Office of Head Start, 2015). The following description pertains to the assessment of the child outcome included in this dissertation.

Devereaux Early Childhood Assessment

A measure of social-emotional development that aligns with the Head Start ELOF, The Devereux Early Childhood Assessment (DECA), is the primary outcome measure used in this dissertation (LeBuffe & Naglieri, 1999; LeBuffe & Naglieri, 2012). DECA is a standardized behavior rating scale completed by caregivers and/or teachers to assess social-emotional protective factors and behavioral concerns in young children (Chain et al., 2010; LeBuffe & Naglieri, 1999; LeBuffe & Naglieri, 2012). The assessment consists of two subscales: Total

Protective Factors, which measures initiative, self-regulation, and attachment/relationships (27 items), and Behavioral Concerns (11 items; LeBuffe & Naglieri, 1999; LeBuffe & Naglieri, 2012).

DECA is a strengths-based measure that evaluates and supports children's positive social-emotional competencies by emphasizing protective factors, allowing educators and caregivers to build upon children's strengths to foster resilience and positive development rather than solely identifying deficits (Bulotsky-Shearer et al., 2013; Chain et al., 2010). While the Behavioral Concerns subscale assesses challenging behaviors, its primary purpose is to contextualize these behaviors within a broader framework of social-emotional development, illustrating how they may influence or interact with the development of protective factors such as *Initiative, Self-Control*, and *Attachment* (Bulotsky-Shearer et al., 2013). Rather than serving as a standalone diagnostic tool, this subscale helps educators and caregivers understand the functional impact of behavioral concerns on a child's ability to build resilience and engage positively in their environment by identifying areas where additional support may be needed (Chain et al., 2010).

Although the Total Protective Factors subscale consistently demonstrates strong construct validity, evidence suggests that the Behavioral Concerns subscale may be better conceptualized as two distinct constructs: internalizing and externalizing behaviors (Bulotsky-Shearer et al., 2013; Chain et al., 2010). This distinction could improve the assessment's ability to identify specific behavioral challenges (Bulotsky-Shearer et al., 2013; Chain et al., 2010). The DECA has been widely used in Head Start programs for early identification of social-emotional needs, yet studies highlight the importance of ongoing validation to ensure its effectiveness across culturally and linguistically diverse populations (Bulotsky-Shearer et al., 2013; Chain et al., 2010). For this dissertation, the change scores from Fall to Spring of DECA Total Protective

Factors and Behavioral Concerns were analyzed as the key child outcomes due to data availability.

Child characteristics may influence DECA scores. For example, pre-K boys tend to exhibit more Behavioral Concerns and lower Total Protective Factors than pre-K girls (Brinkman et al., 2007). However, DECA does not separate norms by gender (LeBuffe & Naglieri, 2012), making it difficult to interpret gender-based differences. Despite this, most DECA items function equivalently across genders, even among children identified as at risk for behavioral problems (Ogg et al., 2010).

Beyond gender differences, children in financially disadvantaged settings are more likely to receive higher Behavioral Concerns ratings and lower Total Protective Factors scores, indicating greater risk for social and developmental challenges (Brinkman et al., 2007). Moreover, regional differences suggest that children in low-income rural Head Start programs tend to score higher in attachment, whereas their urban counterparts score higher in self-regulation, two items of the Total Protective Factors subscale, suggesting that community characteristics influence social-emotional development (Bender et al., 2011). These findings highlight the need to interpret DECA scores within the specific context of Head Start populations to better support children's development.

In summary, the assessment of early childhood outcomes in Head Start programs underscores the importance of measurement tools in understanding and fostering children's development (Office of Head Start, 2015). DECA remains a reliable measure of Total Protective Factors and Behavioral Concerns, consistently demonstrating strong reliability and validity (Bulotsky-Shearer et al., 2013; LeBuffe & Naglieri, 1999; LeBuffe & Naglieri, 2012). However, research emphasizes the importance of contextualizing DECA scores based on child characteristics such as gender, socioeconomic background, and community influences to ensure accurate interpretation and meaningful application (Bender et al., 2011; Brinkman et al., 2007; Ogg et al., 2010). These findings reinforce the ongoing need for validation and refinement of assessment tools to effectively address the unique needs of Head Start populations, particularly in ensuring their applicability across culturally and linguistically diverse groups (Bulotsky-Shearer et al., 2013; LeBuffe & Naglieri, 2012).

CURRENT STUDY

In the current study, I examined the validity and predictive utility of the CLASS Pre-K within Head Start classrooms. The adoption of CLASS Pre-K in the Designation Renewal System reflects an effort to ensure Head Start classrooms maintain high-quality learning environments (Office of Head Start, 2024a). Given that CLASS Pre-K scores determine which programs receive continued funding without competition, its accuracy and reliability are critical. Yet, research conducted since 2020 raises concerns about whether the tool provides a sufficiently robust and valid measure of classroom quality to warrant its high-stakes application in federal funding decisions (Gordon & Peng, 2020; Li et al., 2020; Zheng et al., 2023). Although CLASS Pre-K is intended to provide an objective and reliable assessment of classroom quality, emerging research suggests that its underlying structure and predictive validity remain uncertain (Gordon & Peng, 2020; Hamre et al., 2014; Leyva et al., 2015; Li et al., 2020; Zheng et al., 2023). Despite the widespread adoption of CLASS Pre-K in high-stakes federal funding policy (Office of Head Start, 2024a), there remains a critical gap in research conducted since 2020 that sufficiently validates its predictive utility. Without robust, recent evidence demonstrating its effectiveness in this context, its continued use as a decisive factor in funding decisions may be premature (Gordon & Peng, 2020).

Evaluating the factors that contribute to positive outcomes in Head Start classrooms is essential for maximizing program impact and ensuring effectiveness for diverse populations (Hamre et al., 2014). In this dissertation, I evaluated the validity and reliability of CLASS Pre-K, which measures classroom quality across various domains, referred to as factors in this analysis (Hamre et al., 2014; Pianta et al., 2008b). Additionally, I explore the relationship between

CLASS Pre-K factor scores and child achievement, building on prior research linking classroom quality to positive child outcomes (Cameron et al., 2008; Mashburn et al., 2008).

Research has found modest but statistically significant associations between pre-K center quality and early academic skills (Keys et al., 2013). However, variations in observational assessments raise concerns about the reliability and validity of CLASS Pre-K, particularly in low-income and culturally diverse Head Start settings (Burchinal et al., 2008; Hamre et al., 2014). Given the widespread use of CLASS Pre-K in Head Start and its role in meeting quality benchmarks, it is essential to assess whether its established factor structures remain valid within this context (Gordon & Peng, 2020; Hamre et al., 2014). Ensuring the psychometric robustness of CLASS Pre-K across racially, ethnically, and linguistically diverse populations is essential for accurately measuring classroom quality and its impact on emerging academic skills (Bulotsky-Shearer et al., 2013; Keys et al., 2013).

High-quality early childhood education is particularly vital for addressing the developmental and academic needs of children from disadvantaged backgrounds (Bassok & Galdo, 2016; Burchinal et al., 2008). However, systemic challenges such as limited resources and staffing shortages often hinder the consistent delivery of high-quality instruction, especially in language and literacy development (Pianta et al., 2005; Valentino, 2018). Publicly funded pre-K programs, including Head Start, frequently face these barriers to providing high levels of instructional quality (Bassok & Galdo, 2016; Hamre & Pianta, 2010). Evaluating whether CLASS Pre-K effectively captures variations in classroom quality in these contexts is essential for improving educational outcomes for diverse populations (Hamre et al., 2014; Keys et al., 2013).

In sum, in this dissertation, I evaluated the factor structure of CLASS Pre-K in Head Start classrooms and examined its ability to predict DECA Total Protective Factors and Behavioral Concerns change scores from Fall to Spring. By exploring the applicability of CLASS Pre-K's factor models in a diverse population, I provide valuable insights that can be helpful to program administrators, policymakers, and educators. The goal of this dissertation is to enhance the quality of early childhood education in Head Start programs, thereby promoting equitable opportunities and improving long-term developmental outcomes for children from disadvantaged families. The following research questions were investigated:

1. Is the latent structure of the CLASS Pre-K measure in Head Start classrooms like that found in other populations?
2. Do Head Start classrooms with higher CLASS Pre-K scores yield better improvement in child outcomes from Fall to Spring?

Overview of CLASS Pre-K Predominant Factor Structures

Based on their initial work to develop CLASS Pre-K, Pianta et al. (2008) recommended that the ten CLASS Pre-K dimensions load onto three factors: Emotional Support, Instructional Support, and Classroom Organization. Table 1 presents the recommended three-factor measurement model, which is termed the Teaching Through Interactions Model (Pianta et al., 2008b). The first factor, Emotional Support, includes four dimensions: *Positive Climate, Negative Climate, Teacher Sensitivity*, and *Regard for Child Perspectives*. These dimensions measure the overall emotional support in interactions between teachers and children in the classroom (Pianta et al., 2008b). The second factor, Instructional Support, includes three dimensions: *Concept Development, Quality of Feeback, and Language Modeling*. These dimensions measure the way teachers present concepts that are organized, interconnected, and build upon one another (Pianta et al., 2008b). The third and last factor, Classroom Organization, measures teachers' ability to manage and organize children's time, attention, and behavior in the classroom and includes three dimensions: *Behavior Management, Productivity,* and *Instructional Learning Formats* (Pianta et al., 2008b).

While the three-factor model remains the most widely used in research and practice (Pianta et al., 2008b; Hamre et al., 2013), researchers have explored alternative structures to better capture the complexity of teacher–child interactions (Hamre et al., 2014). The bifactor model was introduced to address the challenge of high correlations among Emotional Support and Classroom Organization, which made it difficult to distinguish their unique contributions to child outcomes (Hamre et al., 2014). By incorporating a general factor, Responsive Teaching, alongside two domain-specific factors, Proactive Management and Routines and Cognitive Facilitation, the bifactor model provides a more precise understanding of how classroom interactions independently support development (Hamre et al., 2014).

Unlike the three-factor model, which treats the ten CLASS Pre-K dimensions as distinct yet interrelated categories, the bifactor model accounts for both overarching and domain-specific influences within a unified structure (Hamre et al., 2014). The general factor, Responsive Teaching, encompasses all ten CLASS Pre-K dimensions, serving as a broad measure of classroom quality (Hamre et al., 2014), while Proactive Management and Routines includes *Positive Climate*, *Negative Climate, Behavior Management*, and *Productivity* to assess classroom structure and organization (Hamre et al., 2014). Cognitive Facilitation consists of *Concept Development*, *Quality of Feedback*, and *Language Modeling*, which together capture the instructional quality of teacher–child interactions (Hamre et al., 2014). By restructuring CLASS Pre-K in this way, the bifactor model mitigates the issue of overlapping domains and strengthens the framework for evaluating classroom interactions (Hamre et al., 2014). Table 2 presents the bifactor measurement model, termed the Effective Teaching Model, which illustrates this revised structure (Hamre et al., 2014).

Prior Factor Analyses of CLASS Pre-K

Numerous studies have examined the factor structure of CLASS Pre-K using one-factor, two-factor, three-factor, and bifactor models. Table 3 includes fit statistics for previous factor analyses of CLASS Pre-K conducted using data at the classroom level (where each classroom has its own row of data with ten columns denoting the scores on each dimension of CLASS Pre-K).While the majority of studies have found support for the three-factor model of CLASS Pre-K (Bihler et al., 2018; Cadima et al., 2018; Downer et al., 2012; Gordon & Peng, 2020; Hu et al., 2016; Karlsen et al., 2023; Leyva et al., 2015; Pakarinen et al., 2010), some researchers have argued that an alternative bifactor model may better capture the complexity of classroom interactions (Hamre et al., 2014; Ng et al., 2021). Since Pianta et al. (2008) published the CLASS Pre-K Manual detailing the predominant three-factor structure, the study to find the strongest support for this structure was completed by Hamre et al. (2013; CFI = .95, RMSEA = .05).

Evidence for the Three-Factor Model

Some factor analyses have found support for the traditional three-factor model of CLASS Pre-K after making modifications, most notably to the dimension of *Negative Climate* (Bihler et al., 2018; Karlsen et al., 2023). Bihler et al. (2018) completed a factor analysis of CLASS Pre-K in state subsidized German preschools and found the three-factor model to fit best after making modifications (CFI = .94, RMSEA = .06, TLI = .95), including dichotomizing the dimension of *Negative Climate* due to low variance. Similarly, Karlsen et al. (2023) investigated the factor structure of CLASS Pre-K among Early Childhood Care and Education groups in Norway and found good fit of the three-factor model (CFI = .93, RMSEA = .07, SRMR = .07, TLI = .92). To achieve their final three-factor model fit, Karlsen et al. (2023) allowed the measurement error of the *Negative Climate* and *Behavior Management* dimensions to correlate, as well as for the dimensions of *Productivity* and *Instructional Learning Formats*. While Bihler et al. (2018) and Karlsen et al. (2023) found that the traditional three-factor model fit best, necessary modifications highlight the complexity of applying CLASS Pre-K across diverse educational contexts and the need for continued examination of its factor structure.

The *Negative Climate* dimension is sometimes excluded entirely from factor analyses due to poor factor loadings and weak correlations with other CLASS Pre-K dimensions (Cadima et al., 2018; Pakarinen et al., 2010). In a factor analysis of CLASS Pre-K using data from pre-Ks in Portugal, Cadima et al. (2018) found that the three-factor model provided the best fit only after modifications, which included removing the *Negative Climate* dimension due to its poor factor loading (CFI = .96, RMSEA = .11, SRMR = .05, TLI = .95). Similarly, Pakarinen et al. (2010) conducted a factor analysis of CLASS Pre-K in Finnish pre-K classrooms and found support for the three-factor model, but their model required removal of the *Negative Climate* dimension and adjustments to residual variances to achieve acceptable fit (CFI = .96, RMSEA = .14, SRMR = .04, TLI = .94). While both analyses found that the three-factor structure of CLASS Pre-K yielded the best fit, necessary modifications suggest that the standard CLASS Pre-K factor structure may not generalize across all contexts (Cadima et al., 2018; Pakarinen et al., 2010). Additionally, the high RMSEA value in Pakarinen et al. (2010; RMSEA = .14) indicates potential model misfit, again reinforcing the need for continued examination of CLASS Pre-K's factor structure across diverse educational settings.

One study retained the dimension of *Negative Climate* in the three-factor model despite a factor loading below the recommended threshold of $\lambda \geq .40$, indicating that this dimension posed challenges for use with their data set (Leyva et al., 2015). Using data from public prekindergarten classrooms in Chile, Leyva et al. (2015) found that the three-factor model fit best (CFI = .93, RMSEA = .11, SRMR = .06, TLI = .88). However, even with retaining *Negative Climate*, modifications were required to achieve adequate fit, including allowing the residuals of eight out of the ten dimensions to correlate (Leyva et al. 2015).

As observed in the aforementioned studies, when investigating the traditional three-factor structure of CLASS Pre-K, modifications are often needed to reach an appropriate fit (Bihler et al., 2018; Cadima et al., 2018; Gordon & Peng, 2020; Karlsen et al., 2023; Leyva et al., 2015; Pakarinen et al., 2010). Most notably, the dimension of *Negative Climate* appears to be a repeated issue, with some studies making modifications to the dimension to keep it in the model (Bihler et al., 2018; Karlsen et al., 2023), others removing the dimension entirely (Cadima et al., 2018; Pakarinen et al., 2010), and some leaving it in even with a poor factor loading (Leyva et al., 2015). For this reason, the *Negative Climate* dimension will require careful consideration in this dissertation.

Evidence for the Bifactor Model

Fewer factor analyses have found support for the bifactor model of CLASS Pre-K (Hamre et al., 2014; Ng et al., 2021). To contribute to the understanding of the CLASS Pre-K factor structure, Hamre et al. (2014) analyzed data from pre-K and Head Start classrooms and found that the bifactor model provided the best fit (CFI = .96, RMSEA = .11, SRMR = .04). In this analysis, the dimension of *Negative Climate* remained in the model despite having a factor loading below the recommended threshold of $\lambda \geq .40$ on the general domain of Responsive

Teaching. Additionally, the dimension of *Language Modeling* had a factor loading of $\lambda < .40$ on the domain-specific factor of Cognitive Facilitation. These findings prompted Hamre et al. (2014) to emphasize the importance of further investigating general and domain-specific approaches to early childhood socialization and education using the bifactor model.

Related, Ng et al. (2021) conducted a factor analysis of CLASS Pre-K using data from kindergarten classrooms (ages 3–6) in Singapore. Ng et al. (2021) reported only Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) fit statistics, making comparisons with other studies challenging due to the less common use of these indices in the field. The bifactor model was found to fit their data best (AIC = 2972.84, BIC = 3098.22; Ng et al., 2021). To achieve this fit, the authors made several modifications: they allowed the dimensions of *Regard for Child Perspectives* and *Instructional Learning Formats* to load only on the general factor of *Responsive Teaching, Negative Climate* to load only on the domain-specific factor of *Proactive Management and Routines*, correlated errors between *Teacher Sensitivity* and *Positive Climate*, and correlated errors between *Productivity* and *Behavior Management.* Notably, *Negative Climate* was the only dimension that did not load onto the general factor of Responsive Teaching. In short, the bifactor model of CLASS Pre-K estimated by Ng et al. (2021) required complex modifications to achieve an appropriate fit.

Evidence for the bifactor model of CLASS Pre-K suggests that extensive modifications are often required to achieve good fit (Hamre et al., 2014; Ng et al., 2021). For example, Hamre et al. (2014) originally proposed a bifactor model with one general factor (Responsive Teaching) and three domain-specific factors (Motivational Supports, Management and Routines, and Cognitive Facilitation). However, Hamre et al.'s (2014) analysis combined Motivational Supports and Management and Routines into a single domain-specific factor, Proactive

Management and Routines, due to high correlations among items. Similarly, Ng et al. (2021) retained the same general factor (Responsive Teaching) and the two domain-specific factors (Proactive Management and Routines and Cognitive Facilitation) as identified by Hamre et al. (2014). In both analyses, the dimension of *Negative Climate* posed challenges, with low factor loadings and issues in model fit (Hamre et al., 2014; Ng et al., 2021). While there is evidence that the bifactor model can provide good fit with certain data sets (Hamre et al., 2014; Ng et al., 2021), the substantial modifications required due to issues with model fit and conceptual complexity of teacher-child interactions, raise concerns about its broader applicability (Ng et al., 2021).

Evaluating Both Models

It has been recommended to evaluate both the three-factor and bifactor models of CLASS Pre-K across diverse populations (Hu et al., 2016). In an analysis of the CLASS Pre-K factor structure among kindergarten classrooms (ages 3–6) in China, Hu et al. (2016) found support for both the bifactor model (CFI = .92, RMSEA = .11, TLI = .92) and the three-factor model (CFI = .91, RMSEA = .11, TLI = .92). In Hu et al. (2016)'s analysis, the dimension of *Negative Climate* demonstrated a good factor loading (λ = -.46) in the traditional three-factor model but a weaker factor loading (λ = .27) on the general factor of Responsive Teaching in the bifactor model. Interestingly, *Negative Climate* obtained a stronger factor loading (λ = -.49) on the domain-specific factor of Proactive Management and Routines (Hu et al., 2016). Hu et al. (2016) recommend considering both the bifactor and the three-factor models of CLASS Pre-K, as model fit can vary depending on the characteristics of the population being assessed. For example, as highlighted in this dissertation, the factor analyses reviewed involve populations from various countries, which may account for inconsistencies in identifying the best-fitting models.

Correlation Matrices Metanalysis

Using an innovative meta-analytic method, Li et al. (2020) utilized Cheung's (2015) metaSEM R package to conduct a two-stage meta-analysis of 26 CLASS correlation matrices (Li et al., 2020). The authors included data from pre-K to 6^{th}-grade classrooms, as these grade levels share the same CLASS dimensions and factors (Li et al., 2020). In stage 1, the matrices were integrated into a single pooled matrix. Homogeneity was assessed using a Q statistic from a multiple-group confirmatory factor analysis (CFA) to determine whether a fixed-effects or random-effects model was appropriate (Li et al., 2020). The results indicated heterogeneity, leading to the adoption of random-effects models to account for individual study-specific matrices (Li et al., 2020). In stage 2, alternative factor structures of CLASS (one-factor, two-factor, three-factor, and bifactor models) were assessed using Weighted Least Squares (WLS) estimation, where correlations with larger sample sizes were weighted more heavily for precision (Li et al., 2020). Among the models evaluated, the bifactor model was discarded due to improper solutions, including negative error variances (Li et al., 2020). Li et al. (2020) found that the two-factor and three-factor models fit the data well, with the three-factor model slightly superior based on fit indices and information criteria[1].

---

[1] Li et al. (2020) found that only 8 out of 91 factor analyses included a CLASS dimension correlation matrix in the publication. After contacting authors, Li et al. (2020) obtained 26 total studies with correlation matrices. The remaining 83 studies provided only factor loadings and between-factor correlations, limiting their usefulness in meta-analytic approaches (Li et al., 2020). This underscores the necessity of including CLASS dimension-level statistics in publications to facilitate meta-analyses and identify broader patterns in the data (Shafer, 2006), such as relationships between CLASS dimensions (Li et al., 2020). The lack of available correlation matrices may contribute to the file drawer problem (Greenwald, 1975; Rosenthal, 1979), where studies with inconclusive or nonsignificant

Cultural Concerns of CLASS Pre-K

Researchers have investigated whether CLASS Pre-K's factor structure, measurement invariance, and predictive validity hold consistently across diverse cultural and demographic contexts (Hamre et al., 2013). Studies across various countries have found support for the three-factor structure, though minor modifications, such as dropping the *Negative Climate* dimension in Finland or allowing correlated errors in Germany, are sometimes necessary for better model fit (Pakarinen et al., 2010; Bihler et al., 2018). In Latin America, Chilean researchers similarly noted that the intended three factors emerged, but allowing certain correlations improved model fit (Leyva et al., 2015). In China, both the three-factor model and a bifactor model provided good fit, suggesting slight differences in how teacher–child interactions are structured in different educational settings (Hu et al., 2016). These findings indicate that while the CLASS Pre-K's core structure is robust, small adaptations may be needed to accurately reflect cultural variations in teaching and interaction styles (Pakarinen et al., 2010; Hamre et al., 2014; Leyva et al., 2015).

To determine whether CLASS Pre-K measures teacher–child interactions equally across diverse classrooms, researchers have tested measurement invariance across different ethnic, linguistic, and socioeconomic groups (Downer et al., 2012). In a study of Latino and dual language learner (DLL) classrooms, CLASS Pre-K's three-factor structure held across these settings, meaning that the instrument measured the same constructs in similar ways regardless of child demographics (Downer et al., 2012). Socioeconomic studies also show that CLASS Pre-K factor scores do not systematically differ based on classroom racial composition, suggesting that

---

findings are less likely to be published (Shafer, 2006). This, in turn, may introduce bias in meta-analyses toward models that appear to receive stronger empirical support, such as the three-factor model of CLASS Pre-K (Li et al., 2020).

CLASS Pre-K is not inherently biased toward rating one type of setting as higher quality than another (Cloney et al., 2017). These findings support the idea that CLASS Pre-K functions reliably across diverse classrooms, although researchers emphasize the need for continued validation in Indigenous and other historically underrepresented settings (Pianta et al., 2008; Downer et al., 2012).

Despite CLASS Pre-K's widespread use, researchers have raised concerns about potential cultural bias in its scoring system (Pastori & Pagani, 2017). One issue is that definitions of high-quality teacher–child interactions vary across cultures, which may influence CLASS Pre-K ratings (Pianta et al., 2008). For example, in Germany, CLASS Pre-K coders reported that American teachers in training videos appeared overly enthusiastic, as such high-energy interactions are not the norm in German preschool culture (Bihler et al., 2018). In other cultural contexts, teachers may place more emphasis on direct instruction and structured learning, which CLASS Pre-K may score lower under *Regard for Child Perspectives* and *Concept Development* (Tan & Rao, 2017). Additionally, CLASS Pre-K does not explicitly measure culturally relevant teaching practices, such as Indigenous language instruction or storytelling traditions, leading some educators to question whether it fully captures diverse approaches to quality education (Cloney et al., 2017). While CLASS Pre-K developers encourage cultural awareness in observer training, critics argue that without explicit adjustments, CLASS Pre-K could unintentionally favor White, middle-class teaching styles over culturally diverse practices (López, 2011; Pastori & Pagani, 2017).

Summary of Prior Factor Analyses

Overall, evidence supports the three-factor model of CLASS Pre-K as generally providing the best fit across diverse data sets. However, the fact that most studies reviewed required modifications to achieve acceptable fit raises concerns about its consistency and applicability across settings (Bihler et al., 2018; Cadima et al., 2018; Hamre et al., 2014; Hu et al., 2016; Karlsen et al., 2023; Leyva et al., 2015; Ng et al., 2021; Pakarinen et al., 2010). Notably, Hamre et al. (2013) was the only study reviewed in this dissertation that did not require any adjustments to reach a solution where all factor loadings exceeded .40, suggesting that the standard model may not generalize equally well across diverse classrooms. In contrast, most studies supporting the three-factor model required adjustments to factor structures, particularly modifications to or removal of the Negative Climate dimension, to achieve an acceptable fit (Bihler et al., 2018; Cadima et al., 2018; Gordon & Peng, 2020; Karlsen et al., 2023; Leyva et al., 2015; Pakarinen et al., 2010). The need for such modifications raises important questions about the validity of CLASS Pre-K, particularly whether it measures teacher–child interactions consistently across different cultural and linguistic contexts or if its structure is inadvertently shaped by the populations in which it was originally developed (Bihler et al., 2018; Cadima et al., 2018; Gordon & Peng, 2020).

Despite these challenges, CLASS Pre-K remains a widely used and influential tool for assessing classroom quality, particularly in high-stakes settings such as the Head Start Designation Renewal System (DRS), where it plays a key role in evaluating program performance (Department of Health and Human Services, 2020). Given the implications of its use in determining Head Start centers' continued funding and operation, it is critical to ensure that CLASS Pre-K accurately reflects classroom quality across diverse populations (Bihler et al., 2018; Cadima et al., 2018; Gordon & Peng, 2020; Karlsen et al., 2023; Leyva et al., 2015; Pakarinen et al., 2010). If modifications are consistently required across studies, this suggests that CLASS Pre-K may need further refinement or cultural adaptations to maintain validity in all early childhood settings (Cadima et al., 2018; Gordon & Peng, 2020). Therefore, continued research on the factor structure of CLASS Pre-K within Head Start classrooms is essential to justify its use in the DRS and to confirm that it provides an equitable and reliable measure of classroom quality across different demographic and cultural contexts (Gordon & Peng, 2020).

The Relationship Between CLASS Pre-K Factors and Outcomes

Table 4 summarizes variables from previous studies that used factor scores from CLASS Pre-K as predictors of improvement in child outcomes from Fall to Spring. While much research has focused on identifying the optimal factor structure of CLASS Pre-K, an equally important question is whether these factors meaningfully predict child outcomes (Hamre et al., 2014; Mashburn et al., 2008). If CLASS Pre-K is to be used as a tool for evaluating classroom quality, its factor scores should reliably correlate with gains in children's academic and social-emotional development (Curby et al., 2009; Sabol et al., 2013). The following section reviews empirical studies that have examined these relationships between CLASS Pre-K and various child outcomes.

Three-Factor Model's Ability to Predict Child Outcomes

The factor of Emotional Support is commonly associated with positive gains in child social competence (Burchinal et al., 2010; Curby et al., 2009a; Downer et al., 2012; Mashburn et al., 2008). Additionally, Emotional Support has been linked to lower behavior problems (Burchinal et al., 2010; Mashburn et al., 2008) and higher executive functioning, including inhibitory control and cognitive flexibility (Leyva et al., 2015). In addition to social-emotional outcomes, Emotional Support has been associated with academic skills such as phonological awareness (Curby et al., 2009b; Hatfield et al., 2016), letter naming (Downer et al., 2012), and language development (Mashburn et al., 2008).

Beyond Emotional Support's role in predicting social emotional outcomes, the factor of Instructional Support is most frequently associated with gains in child academic skills, particularly in language development (Burchinal et al., 2010; Downer et al., 2012; Mashburn et al., 2008), math (Burchinal et al., 2010; Downer et al., 2012), and literacy (Burchinal et al., 2010; Mashburn et al., 2008). It has also been linked to letter naming (Downer et al., 2012) and letter-word identification, though findings can be small and inconsistent (Gordon & Peng, 2020). While Instructional Support primarily influences academic outcomes, some research suggests it may also contribute to closer teacher-child relationships (Hamre et al., 2014).

While Instructional Support primarily shapes academic skills (Burchinal et al., 2010; Downer et al., 2012; Mashburn et al., 2008), Classroom Organization contributes to a broader range of child outcomes, spanning both social-emotional and cognitive domains (Downer et al., 2012; Leyva et al., 2015; Rimm-Kaufman et al., 2009; Hamre et al., 2014). Classroom Organization is associated with a variety of positive child outcomes across academic and social-emotional domains. Academically, it is positively linked to math skills (Downer et al., 2012; Leyva et al., 2015), letter naming (Downer et al., 2012), and literacy (Hamre et al., 2014). In terms of social-emotional skills, Classroom Organization is associated with improved social competence (Downer et al., 2012), behavioral and cognitive self-control, positive work habits and engagement in learning (Rimm-Kaufman et al., 2009). Overall, evidence indicates that the factor of Classroom Organization has a broadly positive impact on both social-emotional and academic skills for children in prekindergarten (Downer et al., 2012; Leyva et al., 2015; Rimm-Kaufman et al., 2009; Hamre et al., 2014).

Bifactor Model's Ability to Predict Child Outcomes

There are fewer outcome studies utilizing the bifactor model of CLASS Pre-K compared to the traditional three-factor model. The findings from Gordon & Peng (2020) and Hamre et al. (2014) offer contrasting perspectives on the predictive utility of the bifactor model. Gordon & Peng (2020) found that the bifactor model had small and generally non-significant associations with gains in children's academic and social skills, suggesting that its factor scores may have limited practical value in predicting child outcomes. Gordan & Peng (2020)'s analysis, based on nationally representative Head Start data, raises concerns about the usefulness of the bifactor model as a reliable indicator of classroom quality in these settings. In contrast, Hamre et al. (2014) reported that the general factor, Responsive Teaching, significantly predicted multiple child outcomes, including language, literacy, and working memory development, as well as reductions in teacher-child conflict among a sample of community-based pre-K classrooms and Head Start programs. Furthermore, the findings in Hamre et al. (2014) suggested that the domain-specific factors within the bifactor model were also meaningful predictors of specific developmental domains. Specifically, Proactive Management and Routines was associated with improvements in inhibitory control, while Cognitive Facilitation predicted gains in early language and literacy skills (Hamre et al., 2014). While Hamre et al. (2014) provides evidence that the bifactor model captures meaningful variations in teacher-child interactions, Gordon & Peng (2020) argue that its limited predictive power, particularly in Head Start classrooms, calls for further refinements or alternative approaches to measuring classroom quality. These inconsistencies highlight the need for additional research to determine whether the bifactor model can be reliably applied across diverse early childhood education settings or whether its structure requires further conceptual and empirical modifications.

Hypotheses

While previous research has established links between CLASS Pre-K scores and child outcomes, the generalizability of its factor structure and predictive validity in Head Start classrooms remains uncertain (Gordon & Peng, 2020). In this dissertation, I address these gaps by evaluating the factor structure of CLASS Pre-K in a diverse Head Start sample and examining whether its factors meaningfully predict social-emotional development. Given its widespread use in assessing classroom quality, concerns persist regarding its applicability across racially, ethnically, and linguistically diverse populations (Downer et al., 2012; Cloney et al., 2017; LoCasale-Crouch et al., 2007; López, 2011). While prior research has found modest but significant associations between pre-K quality and early academic skills, variations in observational assessments raise questions about CLASS Pre-K's effectiveness in capturing meaningful differences in instructional quality (Keys et al., 2013; Mashburn et al., 2008). The findings from this dissertation will provide insight into whether CLASS Pre-K is a reliable measure of classroom quality in Head Start settings. The hypotheses are as follows:

1. The three-factor structure of CLASS Pre-K will demonstrate good model fit in Head Start classrooms, as assessed by established fit indices.
2. CLASS Pre-K factor scores will significantly predict Fall-to-Spring changes in DECA Total Protective Factors and Behavioral Concerns, with higher factor scores associated with positive developmental gains.

METHOD

Participants

The participants in this study included 1,418 children, aged three to five years, and 86 lead teachers from 86 Head Start classrooms across 22 centers in the United States during the 2018–2019 school year. The data set was obtained from the Educare Learning Network, which partners with early childhood centers nationwide to support practice, policy, and research initiatives (Educare, 2024). On average, each classroom had 15 to 16 children. In this sample, 48% of the children were female, 34% were non-Hispanic/Latine white, and 21% were DLLs. All children enrolled in these Head Start classrooms met federal eligibility requirements, which prioritize children from families with incomes at or below the federal poverty level, as well as those experiencing homelessness, children in foster care, and those receiving public assistance such as TANF or SSI (Office of Head Start, 2023).

Only the 86 lead teachers were included in the analyses of the CLASS Pre-K, as the observations for CLASS Pre-K scores focused exclusively on lead teachers and did not include assistant teachers (Office of Head Start, 2023). In this dataset, the average teacher age across all centers was 39 years, 53% of teachers identified as belonging to a historically marginalized race/ethnicity, and over 98% identified as female. Classroom-level teacher data, such as experience, education, or demographics, was excluded from the analysis due to privacy concerns, as Educare noted that divulging this information could lead to deductive identification of the school, classroom, and/or individual.

Sampling

The data for this dissertation was collected during the 2018–2019 school year. The sample of classrooms selected for CLASS Pre-K observations was randomly drawn from the Head Start Enterprise System (HSES), which monitors data for all Head Start centers (Office of Head Start, 2023). A sample of classrooms was drawn from each center. This methodology ensures that the sample provides enough classroom scores representative of each center, as center-level scores are used in the DRS evaluation process (Office of Head Start, 2023).

Procedures

CLASS Pre-K Observations

Data for CLASS Pre-K were collected by reviewers trained and certified through Teachstone, Inc. (Office of Head Start, 2023). After certification, reviewers received ongoing professional development to ensure quality (Office of Head Start, 2023). CLASS Pre-K observations were completed independently by two reviewers, each conducting two 20-minute observation cycles per classroom (Office of Head Start, 2023). Reviewers entered their scores for the ten CLASS Pre-K dimensions into the Head Start Enterprise System (HSES) following the observations (Office of Head Start, 2023). The Office of Head Start Monitoring Software (OHSMS) then averaged the dimension scores from all classrooms within a center to calculate center-level scores for use in the Designation Renewal System (DRS; Office of Head Start, 2023; Department of Health and Human Services, 2020). CLASS Pre-K reviewers did not have access to the center-level scores at any time (Office of Head Start, 2023).

DECA Data Collection

Children in the 86 Head Start classrooms in this dataset were assessed using the Devereux Early Childhood Assessment (DECA; Office of Head Start, 2023). The DECA was completed by teachers who had known each child for at least four weeks (Devereux Advanced

Behavioral Health Center for Resilient Children, 2024; LeBuffe & Naglieri, 2012). Items on the DECA have response options ranging from 0 (never) to 4 (very frequently; Devereux Advanced Behavioral Health Center for Resilient Children, 2024).

Total Protective Factors is a composite measure from the DECA, combining three subscales: Initiative, Self-Regulation, and Attachment/Relationships (LeBuffe & Naglieri, 2012). This composite captures a child's ability to form relationships, manage emotions, and demonstrate independence in daily classroom functioning (LeBuffe & Naglieri, 2012). The Total Protective Factors scale includes 27 items, some of which overlap across the three subscales: Attachment/Relationships (12 items, 6 overlapping), Initiative (16 items, 11 overlapping), and Self-Regulation (11 items, 6 overlapping), with overlapping items counted in more than one subscale (Devereux Center for Resilient Children, n.d.). This composite scale has demonstrated strong internal consistency, with Cronbach’s alpha values ranging from 0.94 to 0.95 for teacher ratings in nationally representative samples (Devereux Center for Resilient Children, n.d.; Lien & Carlson, 2009). Convergent validity has also been supported by positive correlations between Total Protective Factors and other measures of social-emotional competence (Lien & Carlson, 2009).

The Behavioral Concerns subscale consists of 10 items that assess the frequency of externalizing and internalizing behaviors that may interfere with a child’s classroom functioning (Devereux Center for Resilient Children, n.d.). This subscale also demonstrates strong internal reliability, with Cronbach’s alpha ranging from 0.80 to 0.84 in teacher report samples (Devereux Center for Resilient Children, n.d.; Lien & Carlson, 2009). In terms of construct and predictive validity, higher scores on the Behavioral Concerns scale have been associated with increased referrals for behavioral intervention and special education services, supporting its practical relevance in early identification (Lien & Carlson, 2009).

Factor scores (Total Protective Factors and Behavioral Concerns) on the DECA categorize children into three levels—below average, at average, or above average—to help identify strengths and areas of concern in their social-emotional development (LeBuffe & Naglieri, 2012). Children who score at or above average on Total Protective Factors demonstrate age-appropriate or strong skills in attachment, self-regulation, and initiative, which are critical for building resilience and school readiness (LeBuffe & Naglieri, 2012). Conversely, children scoring below average may exhibit delays or difficulties in these areas, indicating a potential need for targeted interventions to support their development (LeBuffe & Naglieri, 2012).

For Behavioral Concerns, higher scores suggest a greater frequency of behavioral challenges, such as difficulty following directions, impulsivity, or struggles with emotional regulation (LeBuffe & Naglieri, 2012). Children with below-average scores in Behavioral Concerns exhibit fewer problem behaviors, while those at or above average may require additional support to develop positive behavioral and coping strategies (Devereux Advanced Behavioral Health Center for Resilient Children, 2024). These classifications help educators and caregivers tailor classroom strategies and interventions to foster positive social-emotional growth in Head Start settings (LeBuffe & Naglieri, 1999).

Before administering the DECA, teachers completed five reflective checklists in both Fall and Spring to evaluate program elements such as the classroom environment, supportive interactions, daily program activities, experiences, and partnerships with families (Devereux Advanced Behavioral Health Center for Resilient Children, 2024). Although these checklists were not recorded, they were intended to help teachers better understand children's behavior and the degree to which the classroom environment supports development (Devereux Advanced Behavioral Health Center for Resilient Children, 2024). Teachers also conducted informal observations of each child during various times of day, settings, and groups over a four-week period to familiarize themselves with the child before completing the DECA (Devereux Advanced Behavioral Health Center for Resilient Children, 2024).

The DECA aligns with the Head Start Early Learning Outcomes Framework (ELOF) domain of social and emotional development, which outlines skills pre-K children need for school success (Devereux Advanced Behavioral Health Center for Resilient Children, 2024; Office of Head Start, 2020). Teachers use these scores at both the child and classroom levels to inform planning and program improvements (Devereux Advanced Behavioral Health Center for Resilient Children, 2024).

DECA’s Psychometric Properties.

The Devereux Early Childhood Assessment (DECA) was first introduced in 1999 and later revised with the release of the DECA-P2 in 2012 (LeBuffe & Naglieri, 1999; LeBuffe & Naglieri, 2012). Although the DECA-P2 aimed to improve upon the original version, its standardization was based on data collected from a sample that may now be outdated, raising questions about its current relevance (LeBuffe & Naglieri, 2012). The demographic characteristics of children and families have shifted over the past decade, suggesting that an assessment tool developed and normed over ten years ago may not adequately reflect the experiences or developmental norms of today’s diverse early childhood population (LeBuffe & Naglieri, 2012).

In addition to concerns about outdated standardization, the DECA’s psychometric properties have also been subject to critique. DECA's factor structure has been reported to show strong internal consistency reliability but found limited discriminant validity among the three subscales—initiative, self-control, and attachment (Barbu et al., 2013). Specifically, they observed high intercorrelations among these constructs, suggesting that the DECA may not adequately differentiate between distinct dimensions of social-emotional functioning (Barbu et al., 2013). This overlap introduces potential ambiguity when interpreting subscale scores and limits the tool's utility for identifying targeted areas of need (Barbu et al., 2013). Together, the age of the instrument and its psychometric limitations suggest a need for updated validation work and possible revision to ensure its continued relevance and precision in contemporary early childhood settings (Barbu et al., 2013; LeBuffe & Naglieri, 2012).

Measures

CLASS Pre-K

For this dissertation, each Head Start classroom's CLASS Pre-K dimension scores collected in Spring average across reviewers were utilized. CLASS Pre-K measures three broad domains, or factors, of teacher-child interactions based on ten dimensions that are critical for children's academic and social-emotional development: Emotional Support, which captures how well teachers establish a warm, supportive classroom climate; Classroom Organization, which assesses the effectiveness of classroom management and productivity; and Instructional Support, which evaluates the ways in which teachers promote higher-order thinking, language development, and concept engagement (Pianta et al., 2008b). Each factor is scored on a 7-point Likert scale based on multiple observation cycles conducted by trained raters (Pianta et al., 2008b).

DECA

For this dissertation, the DECA scores of Total Protective Factors and Behavioral Concerns recorded by teachers for each child were used separately to assess social-emotional development. These outcome scores for each child were transformed into change scores from Fall to Spring. Figure 1 includes density plots of the Fall, Spring, and change scores for both Total Protective Factors and Behavioral Concerns.

Ultimately, two change scores were created by subtracting the Fall score from the Spring score of Total Protective Factors and Behavioral Concerns. A positive change score for Total Protective Factors would indicate a child improved in their ability to think independently, solve problems, take initiative, manage emotions, and form stronger relationships from Fall to Spring, while a negative change score would indicate the child showed a decline from Fall to Spring in these areas (LeBuffe & Naglieri, 1999). A positive change score for Behavioral Concerns would indicate a child displayed more problem behaviors like aggression, withdrawal, and impulsivity from Fall to Spring, while a negative change score would indicate a child displayed fewer of these concerns (LeBuffe & Naglieri, 1999). Additionally, two variables indicating the classroom averages of Total Protective Factors and Behavioral Concerns were created. Table 5 provides descriptive statistics for all variables considered in this analysis.

Data Analysis

Confirmatory Factor Analyses of CLASS Pre-K

Confirmatory factor analyses (CFAs) were conducted to examine the factor structure of CLASS Pre-K within Head Start classrooms, ensuring that the theoretical domains of the instrument align with empirical data. CLASS Pre-K is widely used to assess teacher-child interactions, which are critical for children's developmental outcomes (Pianta et al., 2008b). The data frame used for the CFAs included one row of data per classroom, with ten columns representing the scores on each dimension of CLASS Pre-K. Research on the factor structure of CLASS Pre-K suggests that the widely accepted three-factor model, comprising Emotional Support, Classroom Organization, and Instructional Support, demonstrates acceptable fit across various early childhood settings (Hamre et al., 2014). However, some studies indicate alternative factor structures, particularly in low-income and Head Start populations, where the hierarchical nature of the construct and measurement variations may emerge (Cash et al., 2012; Hamre et al., 2014; Sandilos et al., 2015).

Head Start classrooms serve a diverse population of children, often from economically disadvantaged backgrounds, making it essential to validate the factorial validity of CLASS Pre-K within this setting (Burchinal et al., 2010). Some findings suggest that while the three-factor model holds, modifications may be needed to account for contextual differences in teacher-child interactions specific to Head Start settings (Burchinal et al., 2010; Hamre et al., 2013). Furthermore, measurement invariance testing is often conducted to ensure that CLASS Pre-K functions similarly across different subgroups, including classrooms with varying levels of teacher education, classroom composition, and linguistic diversity (LoCasale-Crouch et al., 2007). Understanding the factor structure within Head Start settings is crucial for ensuring the tool's effectiveness in assessing quality and guiding professional development efforts (Hamre et al., 2013).

Convergent Validity.

Convergent validity examines whether variables that are theoretically related demonstrate strong correlations (Russell, 1978), whereas divergent validity ensures that distinct constructs remain sufficiently independent (Farrell & Rudd, 2009). This step is essential for confirming the conceptual structure of CLASS Pre-K dimensions before proceeding with factor analysis. The evaluation of these validity measures provides a foundation for the subsequent modeling, ensuring that the latent constructs are meaningfully represented by the observed variables (Kline, 2016).

Convergent validity, a key component of construct validity, refers to the extent to which two measures of a theoretically related construct are, in fact, correlated (Russell, 1978). In the context of factor analysis, convergent validity is demonstrated when items hypothesized to measure the same underlying construct (or factor) exhibit strong correlations with one another (Carlson & Herdman, 2012; Russell, 1978). A high degree of correlation among items within the same factor supports the convergent validity of the scale, confirming that these items collectively measure the intended construct (Carlson & Herdman, 2012; Russell, 1978).

Discriminant Validity.

Discriminant validity refers to a latent variable's ability to distinguish itself from other latent variables by accounting for more variance in the observed variables than measurement error, unobserved influences, or other constructs within the same measurement tool (Farrell & Rudd, 2009). In the applied literature, there are multiple definitions and methods for assessing discriminant validity (Rönkkö & Cho, 2022). One widely used approach involves examining the correlations between observed variables, forming a key component of the multitrait-multimethod (MTMM) matrix analysis (Bagozzi et al., 1981; Bagozzi & Phillips, 1982; Campbell & Fiske, 1959). This approach assumes that a measure should *not* strongly correlate with measures it is supposed to differ from (Campbell, 1960). In other words, observed variables associated with different latent constructs should demonstrate minimal correlation.

Correlation coefficients are commonly interpreted using established thresholds: values of $r \leq .35$ indicate weak correlations, $r = .36$ to .67 represent modest correlations, $r > .67$ signify high correlations, and $r \geq .90$ denote extraordinarily strong correlations (Taylor, 1990). Discriminant validity is confirmed when correlations between measures of distinct constructs remain low, ensuring that each latent construct captures a unique aspect of the phenomenon under investigation (Rönkkö & Cho, 2022).

Model Specification.

The three-factor and bifactor models of CLASS Pre-K were included in the analysis for this dissertation. The three-factor model included three factors comprised of ten dimensions: the dimensions of *Positive Climate, Negative Climate, Teacher Sensitivity,* and *Regard for Child Perspectives* loaded onto the factor of Emotional Support, *Behavior Management, Productivity,* and *Instructional Learning Formats* loaded onto Instructional Support, and *Concept Development, Quality of Feedback,* and *Language Modeling* loaded onto Classroom Organization (Pianta et al., 2008b). In this model, the variance standardization method was used which ensures all latent variables are standardized (Kline, 2016). The factor loadings, covariances, and residual variances were also standardized.

The bifactor model included one general factor in which all ten dimensions loaded, as well as two domain specific factors: the dimensions of *Positive Climate, Negative Climate, Behavior Management,* and *Productivity* loaded onto the factor of Proactive Management and Routines, and *Concept Development, Quality of Feedback,* and *Language Modeling* loaded onto Cognitive Facilitation (Hamre et al., 2014). This model was set to be orthogonal, which means the factors are uncorrelated, as is a requirement for bifactor models. Additionally, the variance standardization method and standardized output were used (Kline, 2016).

Measures of Model Fit.

To investigate model fit for the CFAs, the Comparative Fit Index (CFI > .95 representing good fit; Bentler, 1990; Hu & Bentler, 1999), the Root Mean Square Error of Approximation (RMSEA < .06 representing good fit; Browne & Cudeck, 1993; Steiger, 1990), and the Standardized Root Mean Square Residual (SRMR < .08 considered good fit; Hu & Bentler, 1999) were utilized. In addition, factor loadings were examined based on the guidelines of Brown (2015), which suggest that a loading of $\lambda > .40$ typically indicates that an item loads appropriately onto a factor.

CLASS Pre-K Factor Scores as Predictors

The factor scores derived from the CFAs of the three-factor and bifactor models of CLASS Pre-K were then extracted. Several methods are available for deriving scores from the three-factor and bifactor models, as identified by Widaman and Revelle (2023). These methods include: (1) calculating the sum or mean scores of items within each factor, a simple approach that may not account for item-specific variability; (2) exporting factor scores directly from the lavaan package, which computes scores based on each item's contribution to the latent factor; and (3) incorporating confirmatory factor analysis (CFA) into a structural equation modeling (SEM) framework, allowing for the inclusion of confounders and outcomes, thus enabling hypothesis testing within a more complex modeling environment (Widaman & Revelle, 2023).

The third approach was not pursued for two key reasons. First, lavaan (Rosseel, 2012), the software used in this study, has limited capabilities for handling multilevel models, which are necessary given the nested structure of the data (i.e., children within classrooms; Rosseel, 2012). Second, the study aims to evaluate both univariate and multivariate impacts of CLASS Pre-K factors on improvements in child outcomes from Fall to Spring. Univariate effects, which assess the impact of each teaching factor individually, cannot be effectively isolated within a traditional CFA/SEM framework, as this method estimates effects simultaneously and is more suited for multivariate analyses.

When comparing sum/mean scores to factor loading-based methods, each approach has distinct advantages and limitations. Sum/mean scores are straightforward to compute and interpret, making them accessible for researchers and useful for cross-study comparisons (Widaman & Revelle, 2023). They are particularly effective when data are complete and item measurement properties are consistent (Widaman & Revelle, 2023). However, this method assumes all items contribute equally to the construct, potentially oversimplifying relationships and ignoring differential item weightings (Widaman & Revelle, 2023). Conversely, factor loading-based methods provide a nuanced understanding of item contributions to the underlying construct by incorporating weighted contributions based on item importance (Widaman & Revelle, 2023).

For this dissertation, factor estimation was used to calculate factor scores for both the three-factor and bifactor models. These scores were then merged with the rest of the dataset including one row of data per child, with all children in the same classroom sharing the same CLASS Pre-K factor scores to examine the impact of CLASS Pre-K factors on child outcomes.

Multilevel Models

Given the hierarchical nature of the data—where children (Level 1) are nested within classrooms (Level 2), multilevel modeling was the appropriate analytic strategy for examining whether CLASS Pre-K factor scores predict change in child outcomes within Head Start settings (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Ignoring this nested structure would violate the assumption of independent observations required in traditional regression models, potentially leading to biased standard errors and an increased risk of Type I errors, which means incorrectly concluding that CLASS Pre-K predicts child outcomes when it does not (Hox et al., 2018; Luke, 2004). Multilevel models address this by partitioning variance across levels, allowing researchers to separate individual-level effects from contextual or classroom-level influences (Gelman & Hill, 2007; Snijders & Bosker, 2012). This is especially important in early education programs like Head Start, where classrooms vary widely in quality, instructional approaches, and available resources—factors that directly impact developmental outcomes (Burchinal et al., 2010; Downer et al., 2012).

In this dissertation, the two outcome variables—change scores from Fall to Spring for DECA Total Protective Factors and Behavioral Concerns—were regressed on the CLASS Pre-K standardized factor scores from both the traditional three-factor model and the bifactor model. Multilevel models with random intercepts were used to account for classroom-level variability in child growth trajectories (Koth et al., 2008; McCoach et al., 2013). To further control for classroom context, each model included the classroom average of the outcome variable at Fall as a covariate, capturing the influence of being in a higher- or lower-functioning classroom on individual children's developmental progress (Enders & Tofighi, 2007; Wang & Maxwell, 2015).

Due to the small number of Head Start centers in the sample, a fixed effects approach was used to account for third-level clustering where classrooms are nested within centers (Raudenbush & Bryk, 2002). This was implemented through the inclusion of $k-1$ dummy variables, where *k* represents the number of centers, to control for unobserved between-center differences without assuming normally distributed random effects (Bell et al., 2010). This strategy allows researchers to isolate classroom-level effects more precisely, particularly when center-level variance cannot be robustly estimated due to sample size limitations (Bell et al., 2010; Raudenbush & Bryk, 2002).

Model estimation was conducted using the lmer function from the *lme4* package in R, utilizing the Restricted Maximum Likelihood (REML) method, which is appropriate for estimating variance components in multilevel models (Bates et al., 2015). This analytic framework aligns with best practices in early childhood education research, where both teacher-level factors and child-level differences must be accounted for to accurately evaluate the effects of classroom quality on developmental outcomes (Pianta et al., 2008; Mashburn et al., 2008; Burchinal et al., 2010). Figure 2 depicts the hypothesized associations between CLASS Pre-K factor scores and child change scores in DECA Total Protective Factors and Behavioral Concerns from Fall to Spring.

Model Specification: Three-Factor Model.

With the three-factor model, it is necessary to address collinearity, or multicollinearity, which refers to the condition where at least two predictor variables in a statistical model are highly linearly related, creating challenges for statistical estimation and interpretation (Dormann et al., 2012). This interdependence makes it difficult to isolate the effects of individual predictors, inflates standard errors, and leads to biased parameter estimates, which can compromise the reliability of inferences (Dormann et al., 2012). A commonly recognized threshold for problematic collinearity is when the correlation coefficient between predictors exceeds $r = .70$ (Dormann et al., 2012). In this dataset, the Emotional Support and Classroom Organization domains of CLASS Pre-K are highly correlated ($r = .96$), making it challenging to disentangle their individual effects. Removing one domain would result in partial loss of explanatory power, while including both risks instability in model predictions due to collinearity

(Dormann et al., 2012). Additionally, high collinearity may cause small variations in the dataset, such as minor changes in values or the inclusion or exclusion of a few observations, to disproportionately affect model outcomes by making parameter estimates unstable and inflating standard errors, which can lead to significant fluctuations in predictions and statistical significance (Dormann et al., 2012).

To address these issues related to multicollinearity, a sequence of four models for the three-factor model of CLASS Pre-K was implemented for each outcome. The first three models for each outcome examined the effect of each classroom quality factor (Emotional Support, Instructional Support, and Classroom Organization) individually, allowing for the evaluation of their unique contributions while accounting for the multilevel data structure. The fourth model for each outcome incorporated all three domains simultaneously, assessing their independent effects while controlling for overlap. This approach aimed to mitigate the impact of collinearity and ensure robust and interpretable results. The individual coefficients for each model were evaluated.

Model Specification: Bifactor Model.

Similarly to the three-factor model, separate models for each of the two outcomes were fit for the bifactor model analysis. The first model for each outcome included only the general factor, termed Responsive Teaching, to gauge its overall impact. The second model for each outcome was extended to include the general factor alongside the two specific factors: Proactive Management and Routines, and Cognitive Facilitation. This model aimed to understand both the combined and individual contributions of these factors to the improvement in child outcomes from Fall to Spring. As with the three-factor model, the individual coefficients, and the proportional reduction in error for each model were evaluated.

Outcomes and Missing Data.

Because model estimation depends on the available outcome data, it is important to first consider the extent of missingness in the outcome variables. Of the 1,655 children originally included in the dataset for this dissertation, 25 were missing their classroom ID. This issue arose because Local Evaluation Partners, who conduct CLASS Pre-K observations, failed to enter the classroom ID initially and did not respond to error reports issued by the Frank Porter Graham Child Development Institute, which oversees the Educare National Evaluation. As a result, these children lacked an assigned classroom ID and were excluded from the analysis.

Additionally, several centers and numerous classrooms were missing data for CLASS Pre-K, the study's primary predictor. This occurred because not all centers were required to conduct CLASS Pre-K observations for every classroom each year. To ensure consistency, the dataset was restricted to 86 classrooms across 19 centers that had completed CLASS Pre-K assessments. After removing classrooms without CLASS Pre-K scores, the final sample included 1,418 children.

Lastly, the primary outcomes of this study were change scores derived from the Fall and Spring assessments of two components of the DECA (Devereux Advanced Behavioral Health Center for Resilient Children, 2024; LeBuffe & Naglieri, 1999). However, five of the 19 centers only submitted Fall DECA scores and did not provide Spring assessments, as DECA administration varies across sites.

Examining Selection Bias.

Classrooms that submitted both Fall and Spring DECA scores were compared with those that only submitted Fall scores to assess potential differences. Specifically, variations in CLASS Pre-K scores and Fall DECA scores were examined using a Wilcoxon rank-sum test. While

CLASS Pre-K scores did not significantly differ between the two groups, Fall DECA scores showed significant differences, suggesting potential selection bias.

For the DECA Total Protective Factors score, the mean Fall score for classrooms that submitted Spring DECA data (n = 65) was 48 (SD = 6), whereas classrooms without Spring DECA scores (n = 21) had a higher mean score of 52 (SD = 7), indicating a significant difference (p = .017). Since higher scores on this measure are more desirable, classrooms without Spring scores exhibited stronger baseline protective factors.

Similarly, for the DECA Behavioral Concerns score, classrooms with Spring DECA scores had a mean Fall score of 51 (SD = 4), while those without Spring DECA scores had a lower mean score of 47 (SD = 6), representing a significant difference (p = .035). As lower scores on this measure are more desirable, classrooms without Spring scores had fewer behavioral concerns at baseline.

Multiple Imputation.

Multiple imputation is a statistical method used in this dissertation to address missing outcome data, the DECA Total Protective Factors and Behavioral Concerns change scores, by generating multiple independent datasets with imputed values, capturing the uncertainty surrounding the true values of missing observations (van Buuren, 2018). Instead of substituting a single estimated value, multiple imputation replaces each missing data point with multiple plausible values, creating *m* complete datasets (van Buuren, 2018). These datasets are then analyzed separately, and their results are combined to produce final estimates. This approach is particularly valuable for maintaining statistical power and reducing bias in analyses involving missing data (van Buuren, 2018). In this dissertation, missing data primarily involved child behavioral outcome scores in Fall and Spring, which were subsequently used to calculate change scores. Since these variables were necessary for the multilevel models, multiple imputation was specifically applied within that framework.

To mitigate potential biases arising from differences in Fall scores between classrooms that submitted and did not submit Spring scores, a comprehensive set of auxiliary variables was included in the imputation process. The imputation model incorporated all six CLASS Pre-K predictors (Emotional Support, Instructional Support, Classroom Organization, Responsive Teaching, Proactive Management and Routines, and Cognitive Facilitation), classroom-level averages for Total Protective Factors and Behavioral Concerns in both Fall and Spring, as well as child-level scores for Total Protective Factors and Behavioral Concerns in both Fall and Spring. This ensured that the imputation process leveraged all available information to generate the most accurate estimates.

The multiple Imputation process followed three key steps. First, missing values were imputed using the 2l.norm method from the mice package in R (van Buuren & Groothuis-Oudshoorn, 2011). This method is designed specifically for two-level hierarchical data, where lower-level observations (e.g., children) are nested within higher-level units (e.g., classrooms; van Buuren & Groothuis-Oudshoorn, 2011). Unlike standard imputation techniques that treat data as a single level, 2l.norm preserves the hierarchical structure by modeling both fixed and random effects (van Buuren, 2018). This ensures that the imputed values maintain the natural clustering of the data, reducing bias and improving statistical validity (Enders et al., 2016).

Second, the imputation process employed Gibbs sampling, a Bayesian approach that iteratively estimates missing values by drawing from the posterior distribution of the data (van Buuren, 2018; Kasim & Raudenbush, 1998). Bayesian methods ensure that each imputed value is statistically plausible by generating posterior distributions that account for the variability and structure of the observed data. In cases where classrooms were missing all child-level data, the 2l.norm method relied on the distributions estimated from classrooms with observed data, further improving the accuracy of imputations (Enders et al., 2016).

Finally, after generating the imputed datasets, the multilevel models were fitted separately to each of the *m* complete datasets. The results were then pooled using Rubin's rules (Rubin, 1987, as cited in van Buuren, 2018), which adjust for both within-imputation and between-imputation variability. This pooling process ensures that final estimates incorporate the uncertainty associated with missing data, leading to more robust and reliable inferences (van Buuren, 2018). By implementing multiple imputation with 2l.norm (van Buuren & Groothuis-Oudshoorn, 2011), this dissertation increased the likelihood that potential biases introduced by missingness were properly accounted for while preserving the hierarchical nature of the data. The integration of Bayesian posterior estimation further enhances the validity of the imputed values, ensuring that the resulting estimates accurately reflect the underlying data distribution (van Buuren, 2018; Kasim & Raudenbush, 1998).

Change Scores.

After completing multiple imputation for both the Fall and Spring DECA Total Protective Factors and Behavioral Concerns scores, absolute change scores were calculated. Utilizing change scores in this analysis is critical because they help mitigate bias introduced by unobserved confounders that influence individual children' Fall and Spring scores (Burchinal et al., 2010; Winship & Morgan, 1999). For example, a child who begins with a higher Fall score relative to their peers is more likely to maintain a similarly high score in the Spring. Without accounting for child Fall scores by fitting a model that only includes Spring scores, the interpretation of results could be skewed, potentially leading to spurious significant findings (Winship & Morgan, 1999).

By using change scores, the analysis reduces bias at the child level stemming from factors such as baseline skill level, demographic characteristics, dual language learner status, and disability status. Change scores operate under the assumption that individuals have stable, fixed differences, resulting in regression toward their individual means rather than toward the overall population mean (Winship & Morgan, 1999). This approach enhances the sensitivity of the analysis to meaningful differences in improvement in child outcomes from Fall to Spring, ensuring that the focus remains on changes occurring within individuals rather than on comparisons to the broader classroom. By emphasizing within-individual changes, the analysis can more precisely capture how variations in instructional quality measured by the CLASS Pre-K factors contribute to improvement in child outcomes from Fall to Spring while mitigating the risk of confounding due to broader classroom-level influences which, if unaccounted for, can lead to misleading conclusions about the effects of specific CLASS Pre-K factors (Winship & Morgan, 1999; Burchinal et al., 2010).

Confounders.

To assess the causal relationship between classroom quality and child outcomes, it is essential to identify and account for all relevant confounders in addition to individual-level difference represented by change scores from Fall to Spring (VanderWeele, 2019). A confounder is a variable that influences both the exposure—in this case, CLASS Pre-K scores—and the child outcomes, DECA scores (VanderWeele, 2019). Importantly, these confounders must be causally prior to both CLASS Pre-K scores and child outcomes (VanderWeele, 2019).

Child-level variables alone cannot be considered confounders because they do not have the capacity to influence CLASS Pre-K scores. However, when child-level characteristics are aggregated at the classroom level, they may affect CLASS Pre-K scores (Winship & Morgan, 1999). By incorporating these aggregated measures, potential confounding effects—such as having many disruptive children or a predominance of high-achieving children, which could respectively hinder or enhance child outcomes—are accounted for (Winship & Morgan, 1999). For example, a high concentration of disruptive children might negatively impact the overall classroom learning environment, whereas a classroom with high-achieving children might provide more advanced learning opportunities. To account for these effects, the classroom-level averages of DECA Total Protective Factors and Behavioral Concerns scores based on Fall assessments were computed and incorporated as Level 2 confounders in the multilevel models. Since CLASS Pre-K observations typically occur in the Spring, using Fall-aggregated outcome scores ensures that these measures are causally prior to the CLASS Pre-K factor scores and change scores from Fall to Spring, strengthening the interpretation of their role as confounders.

Additionally, this analysis included a fixed effect for centers by incorporating dummy-coded variables representing the different Head Start centers. Dummy coding is a method for converting categorical variables into numerical representations for regression analysis (Hardy, 1993; Hardy & Reynolds, 2004; Skrivanek, 2009; Wolf & Cartwright, 1974). To control for variability across centers, 18 dummy-coded variables were created to represent the 19 centers. Specifically, a column in the dataset was generated for each center, assigning a value of 0 for classrooms not belonging to that center and 1 for those that did. These dummy variables were included as Level 2 covariates in the multilevel model. This step is crucial because the centers, which are distributed across different regions in the United States, may exhibit significant demographic and contextual differences (Burchinal et al., 2010; Keys et al., 2013). By incorporating center as a fixed effect, the model effectively controlled for systematic differences across locations, such as variations in community demographics or administrative policies, which might otherwise introduce bias into the estimated relationships.

Model Fit.

Model fit was assessed using multiple approaches. First, the significance levels of each factor in both the bifactor and three-factor models were examined. In this context, if the key predictor, such as Emotional Support, has a significant slope, it indicates that for two classrooms differing by one unit in their average Emotional Support score, the outcome measure is expected to increase by the value of the estimated slope (Nezlek, 2012). Conversely, if a key predictor is not statistically significant, it suggests that there is insufficient evidence to conclude that an effect of CLASS Pre-K on child outcome scores exists given the sample size and variability in the data (Nezlek, 2012). Instead, a non-significant result may reflect limited statistical power, measurement error, or the presence of unmeasured confounders (Nezlek, 2012).

Since the outcome change scores were computed as Spring score minus Fall score, a positive slope suggests that higher Emotional Support is associated with greater increases in the outcome over time, whereas a negative slope would indicate that higher Emotional Support is linked to a smaller increase or a decline in the outcome. For example, in the case of Total Protective Factors, a positive coefficient would mean that children in classrooms with higher Emotional Support experienced greater improvements in social-emotional strength and resilience from Fall to Spring. For Behavioral Concerns, which is scored in the opposite direction of Total Protective Factors, a positive coefficient would indicate that children in classrooms with higher Emotional Support experienced an increase in Behavioral Concerns from Fall to Spring, suggesting a decline in behavioral regulation. Conversely, a negative coefficient would mean that higher Emotional Support was associated with a greater reduction in Behavioral Concerns, indicating improved self-regulation and fewer problematic behaviors over time.

R^2 Values.

R^2 values were assessed for between-group variance to evaluate the proportion of variance explained by the model, serving as a measure of effect size (Rights & Sterba, 2019). R^2 values range from 0 to 1, with 0 indicating that no variance is explained and 1 indicating that all variance is explained by the model (Shaw et al., 2022). These values are particularly useful because they allow for comparisons across studies (Rights & Sterba, 2019).

The r2mlm package In R was utilized to compute R^2 effect size measures that quantify the explained variance at distinct levels of the model, including total variance, within-cluster variance, and between-cluster variance (Rights & Sterba, 2019; Shaw et al., 2022). Specifically, in this analysis, R^2 values were calculated for total variance and variance between classrooms. Variance within classrooms was not calculated as there were no predictors at this level added to the model. A key advantage of the r2mlm package is that it automates the computation of multiple R^2 measures and provides graphical representations of these effect sizes, offering a comprehensive view of the model's explanatory power (Shaw et al., 2022).

For this dissertation, R^2 values were calculated separately for each of the data sets derived from multiple imputation, and the final R^2 estimate was obtained by averaging across all imputed data sets. This approach ensures that the reported R^2 values accurately reflect the overall explanatory power of the model while accounting for missing data.

Intraclass Correlations.

To assess the proportion of between-group variance before adding predictors to the model, intraclass correlations (ICCs) were obtained by estimating separate unconditional multilevel models with random intercepts for each outcome (Kuppens & Yzerbyt, 2014). ICCs provide an estimate of the proportion of variance attributable to differences between classrooms relative to the total variance (Kuppens & Yzerbyt, 2014). The ICC was calculated using the standard equation:

$$ICC = \frac{Variance_{classroom}}{Variance_{classroom} + Variance_{student}}$$

This analysis tested whether variance differed significantly between classrooms or whether it was dependent on the continuous measures of CLASS Pre-K factors (Kuppens & Yzerbyt, 2014). Furthermore, this approach allowed for an assessment of whether residual variance was influenced by main effects and interactions, ensuring that the inclusion of predictors accounted for meaningful variation rather than unstructured error (Kuppens & Yzerbyt, 2014).

Power analysis

A power analysis was conducted using the simr package in R (Green & MacLeod, 2016). The first step involved specifying a multilevel model based on observed data parameters, where classroom quality was used as a predictor of change scores. This approach entailed simulating data using an initial model with a candidate classroom quality measure, then adjusting the unstandardized effect size to align with the expected standardized effect size of 0.30. The simulations indicated that to achieve 80% power with an alpha of 0.05 and assuming an average of 14 children per classroom, a minimum of 50 classrooms was required for DECA Behavioral Concerns and 56 classrooms for DECA Protective Factors.

Additionally, further power analysis guidance was obtained from Arend and Schäfer (2019), who provide benchmarks for multilevel models based on intraclass correlation coefficients (ICCs) and level-1 sample sizes. For a medium ICC (0.30) and a level-1 sample size of 14 children per classroom, their findings suggest that 90 classrooms would be necessary to detect a standardized effect size of 0.30 with 80% power.

Both approaches highlight a significant discrepancy between the available data and the recommended sample size for robust power. The dataset includes only 44 classrooms with DECA Total Protective Factors scores and 43 classrooms with DECA Behavioral Concerns scores that have at least 14 children with observed change scores. This discrepancy underscores the challenges posed by missing data and emphasizes the limitations in statistical power due to reduced sample size.

RESULTS

The analysis presented in this dissertation consisted of two components: (1) two confirmatory factor analyses (CFAs) evaluating the three-factor and bifactor models of CLASS Pre-K, and (2) a series of multilevel models examining the relationship between CLASS Pre-K factors and improvement in child outcomes from Fall to Spring. Additionally, the html R notebook including the code and output for this dissertation can be downloaded at this link.

Factor Structure Analysis

Table 6 presents correlation coefficients for the ten CLASS Pre-K dimensions. Additionally, Figure 3 depicts these associations through a scatterplot correlation matrix, allowing for a clearer interpretation of patterns among CLASS Pre-K dimensions.

Convergent Validity

Before fitting the factor models, an initial assessment of convergent and divergent validity was conducted. CLASS Pre-K dimensions within the same factor are expected to exhibit some degree of convergent validity, as they measure related constructs rather than entirely distinct aspects of classroom quality.When defining factors, it is recommended to use measures with a convergent validity correlation of $r > .70$ and to avoid those with $r < .50$ (Carlson & Herdman, 2012). Assessing convergent validity is crucial because even when a measure demonstrates a high degree of convergent validity—for example, a correlation of $r = .85$ with another theoretically related measure—substantial variation in effect sizes across studies can still occur. This variability can complicate the interpretation of findings, as it raises concerns about the consistency and generalizability of the relationship across different samples or contexts (Carlson & Herdman, 2012).

Three-Factor Model.

For the factor of Emotional Support, *Positive Climate, Teacher Sensitivity, and Regard for Child Perspectives*, were correlated with *Negative Climate* at r = -0.43 to r = -0.47. The correlations between *Positive Climate, Teacher Sensitivity, and Regard for Child Perspectives* ranged from r = 0.63 to r = 0.64. These correlations fall below the recommended threshold of $r >$ 0.70 (Carlson & Herdman, 2012), providing some evidence for the potential removal of *Negative Climate* as a dimension of Emotional Support. The negative correlations arise because *Negative Climate* is reverse-scored relative to the other dimensions. However, even among the other dimensions, correlations remained below r = 0.70, indicating that they do not meet the recommended level for convergent validity (Carlson & Herdman, 2012). That said, none of the correlations fell below the threshold for avoidance (r < 0.50; Carlson & Herdman, 2012), suggesting that the Emotional Support factor exhibits marginal convergent validity.

For the factor of Instructional Support, the three dimensions (*Concept Development, Quality of Feedback, and Language Modeling*) were correlated at r = 0.61 to r = 0.74. The only dimension pair that met the recommended threshold of r > 0.70 was *Concept Development* and *Quality of Feedback* (Carlson & Herdman, 2012). *Language Modeling* correlated below r = 0.70 but remained above r = 0.50 with both *Quality of Feedback* and *Concept Development*, meaning that while it does not meet the standard recommendation for convergent validity, it does not fall within the range that suggests elimination (Carlson & Herdman, 2012). The weaker correlations involving *Language Modeling* suggest that the Instructional Support factor also exhibits only marginal convergent validity.

Finally, for the factor of Classroom Organization, the three dimensions (*Behavior Management, Productivity, and Instructional Learning Formats*) were highly correlated at r =

0.73 to *r* = 0.78. These strong correlations indicate a consistent relationship among the three dimensions, where increases in one dimension correspond with increases in the others (Carlson & Herdman, 2012). With all correlations exceeding *r* > 0.70 (Carlson & Herdman, 2012), Classroom Organization is the only factor in the three-factor model of CLASS Pre-K that demonstrates strong convergent validity in this dissertation.

Bifactor Model.

In the bifactor model, the domain-specific factors of Proactive Management and Routines and Cognitive Facilitation are expected to exhibit moderate to high correlations among their corresponding dimensions. Within the Proactive Management and Routines factor, all dimensions demonstrated moderate to high correlations ($r \geq 0.61$), apart from *Negative Climate*, which showed a weaker negative correlation ($r \leq -0.47$*). Similarly, the Cognitive Facilitation factor displayed moderate correlations among its dimensions ($r \geq 0.61$). These results provide moderate evidence of convergent validity for the dimensions within the two domain-specific factors in the bifactor model of CLASS Pre-K.

Discriminant Validity

Three-Factor Model.

In the context of CLASS Pre-K, discriminant validity requires that dimensions from distinct factors exhibit low correlations, indicating that they measure distinct constructs. However, an examination of intercorrelations within the three-factor model suggests potential issues with discriminant validity between the factors of Emotional Support and Classroom Organization. Specifically, *Positive Climate,* a dimension within Emotional Support, showed moderate correlations with *Behavior Management* (*r* = .61), *Productivity* (*r* = .65), and *Instructional Learning Formats* (*r* = .52), all of which belong to Classroom Organization.

Similarly, *Teacher Sensitivity*, also from Emotional Support, demonstrated moderate to high correlations with the Classroom Organization dimensions: *Behavior Management* (r = .80), *Productivity* (r = .68), and *Instructional Learning Formats* (r = .56). *Regard for Child Perspectives* followed the same pattern, correlating moderately to highly with *Behavior Management* (r = .67), *Productivity* (r = .72), *and Instructional Learning Formats* (r = .63). These stronger-than-expected intercorrelations between dimensions of Emotional Support and Classroom Organization raise concerns about the discriminant validity of the three-factor model of CLASS Pre-K.

Bifactor Model.

In bifactor models, domain-specific factors, such as Proactive Management and Routines and Cognitive Facilitation, are intended to measure distinct constructs. As a result, their corresponding dimensions should exhibit low correlations with each other. In this analysis, *Behavior Management* from Proactive Management and Routines and *Language Modeling* from Cognitive Facilitation were the only two dimensions that demonstrated a moderate correlation across these domain-specific factors (r = .50). This finding provides evidence supporting the discriminant validity of the two domain-specific factors in the bifactor model of CLASS Pre-K. With convergent and discriminant validity assessed, the next step was to examine the factor structure.

Confirmatory Factor Analyses

The lavaan package in R (Rosseel, 2012) was used to examine the two most researched models of CLASS Pre-K: the three-factor model (Pianta et al., 2008b) and the bifactor model (Hamre et al., 2014). The details for model specification are provided in the subsequent sections. Factor loadings and variances for both confirmatory factor analysis (CFA) models, along with covariances for the three-factor model, are presented in Table 7. Fit indices for both confirmatory factor analyses are displayed in Table 8.

CFA Model 1: The Three-Factor Model

CFA Model 1 followed the traditional three-factor structure initially proposed by Pianta et al. (2008). This model included three latent factors: Emotional Support, Classroom Organization, and Instructional Support. Factor variances were constrained to 1, while factor loadings were freely estimated, and the factors were allowed to covary. All dimensions in the three-factor model demonstrated factor loadings of $\lambda > .40$, suggesting meaningful relationships with their respective factors (Brown, 2015). The lowest factor loading was observed for *Negative Climate* ($\lambda = -0.53$), which, despite being negative due to its reverse scoring, still exceeded the recommended threshold ($\lambda > .40$). However, model fit indices indicated poor fit (χ^2 (32, n = 86) = 78.07, CFI = .92, TLI = .89, RMSEA = .13, SRMR = .08).

An additional method for assessing discriminant validity within the multitrait-multimethod approach is to examine the correlations between latent constructs (Kenny, 1976; McKenny et al., 2013; Reichardt & Coleman, 1955; Rönkkö & Cho, 2022; Schmitt & Stults, 1986; Werts & Linn, 1970). In confirmatory factor analysis (CFA), the variance standardization method constrains factor variances to 1, thereby producing latent variable correlation coefficients in the model output (Rosseel, 2012). Table 9 presents the correlation matrix for the three-factor model of CLASS Pre-K, facilitating an additional evaluation of discriminant validity within this dataset. The correlation between Classroom Organization and Emotional Support was high (r = .96), suggesting that these two factors function similarly and may not represent distinct constructs (Rönkkö & Cho, 2022). In contrast, Emotional Support and Instructional Support exhibited a low to moderate correlation (r = .47), as did Classroom Organization and

Instructional Support (r = .52; Taylor, 1990). Given these findings, the three-factor model of CLASS Pre-K demonstrates limited discriminant validity within this dataset.

CFA Model 2: The Bifactor Model

A bifactor model includes a general factor that all items load onto, along with multiple domain-specific factors that capture unique item groupings (Dunn & McCray, 2020). In this structure, each item's variance is divided between the general and a specific factor, enabling an assessment of both shared and unique contributions (Dunn & McCray, 2020). These specific factors are assumed to be orthogonal to the general factor, meaning they are uncorrelated (Dunn & McCray, 2020). For CLASS Pre-K, Hamre et al. (2014) proposed a bifactor structure featuring a general factor—Responsive Teaching—alongside two domain-specific factors: Proactive Management and Routines, and Cognitive Facilitation. This modeling approach is especially useful for evaluating how much variance is attributable to specific item clusters after accounting for the overarching construct (Dunn & McCray, 2020). Additionally, bifactor models help determine whether domain-specific subscales offer distinct information beyond the general factor, thereby supporting their construct validity (Reise et al., 2010; Reise et al., 2018). In the context of CLASS Pre-K, this approach tests whether Responsive Teaching can be empirically differentiated from the more targeted domains of Proactive Management and Routines and Cognitive Facilitation.

CFA Model 2 represents the bifactor model of CLASS Pre-K (Hamre et al., 2014). The factor variances were standardized to 1, and factor loadings were freely estimated, with all factors specified as uncorrelated (i.e., covariances were set to zero). This model allows for an assessment of how each dimension relates to its designated factors. Notably, several factor loadings for the general factor of Responsive Teaching and the domain-specific factor of

Proactive Management and Routines fell below the commonly accepted threshold of $\lambda > .40$ (Brown, 2015). Specifically, *Concept Development*, *Quality of Feedback*, *Negative Climate*, *Behavior Management*, and *Productivity* demonstrated weak loadings. Moreover, *Behavior Management* exhibited a negative loading on Proactive Management and Routines, which contradicts theoretical expectations, as higher scores on *Behavior Management* should align with improvements in Proactive Management and Routines. The overall fit of the bifactor model was marginal, as indicated by the fit statistics (χ^2 (28, n = 86) = 59.93, CFI = .94, TLI = .91, RMSEA = .12, SRMR = .06).

Best Fitting Model

In confirmatory factor analyses (CFAs), chi-square (χ^2) serves as an absolute fit index, where lower values relative to the degrees of freedom indicate better model fit (Alavi et al., 2020). However, chi-square is extremely sensitive to sample size, often leading to significant results even when model misfit is minor. To account for this, researchers frequently use the chi-square statistic divided by the degrees of freedom (χ^2/df) as a relative fit measure. While some researchers have proposed that a ratio of ≤ 2 indicates good fit, and values < 5 may still be acceptable, these thresholds remain debated, as they lack a strong statistical basis (Kline, 2016).

For this study, the ratio between the chi-square statistic and the degrees of freedom was 2.44 for the three-factor model and 2.14 for the bifactor model. These values do not meet the stricter criterion of ≤ 2 but satisfy the more lenient threshold of < 5, indicating an acceptable, though not optimal, model fit.

Factor Loadings.

In the three-factor model of CLASS Pre-K, all dimensions loaded as expected, exceeding the $\lambda > .40$ threshold (Brown, 2015), indicating strong associations between each dimension and its respective factor. In contrast, the bifactor model exhibited several dimensions with loadings below this threshold (Brown, 2015), along with an unexpected negative factor loading. These issues suggest that the bifactor model may not fully capture certain domains of CLASS Pre-K with optimal accuracy.

Fit Statistics.

Despite the appropriate factor loadings, the three-factor model of CLASS Pre-K obtained poor fit statistics (CFI = .92, TLI = .89, RMSEA = .13, SRMR = .08), suggesting that the structure of the three-factor model might not align well with this data set. Moreover, the bifactor model showed marginal fit (CFI = .94, TLI = .91; RMSEA = .12, SRMR = .06). The bifactor model obtained an improvement in fit over the three-factor model, but these fit statistics are still not within range of what is considered excellent fit. Furthermore, the slight difference in fit between the two models is negligible because the bifactor model should mathematically fit better due to the number of regression paths doubling compared to the three-factor model, enhancing its ability to explain variance.

Theoretical Alignment.

One strength of the three-factor model of CLASS Pre-K is its theoretical alignment with the original structure proposed by Pianta et al. (2008). Another advantage is its clear distinction between factors of positive classroom interactions—Emotional Support, Classroom Organization, and Instructional Support. This distinctiveness allows for straightforward interpretation of statistical results. Additionally, this model is well-established in the literature, with several studies investigating fit for the three-factor model of CLASS Pre-K (Bihler et al., 2018; Cadima et al., 2018; Downer et al., 2012; Gordon & Peng, 2020; Hamre et al., 2013; Karlsen et al., 2023; Leyva et al., 2015; Pakarinen et al., 2010). However, a key limitation of the three-factor model is that, while CLASS Pre-K is designed as a comprehensive measure of classroom quality, this structure does not provide an overall classroom quality score. Instead, its implementation requires individuals to infer overall classroom quality based on separate factor scores (Pianta et al., 2008b).

The bifactor model of CLASS Pre-K offers an advantage over the three-factor model by incorporating a general classroom quality factor, Responsive Teaching, which allows for more nuanced analyses (Dunn & McCray, 2020; Hamre et al., 2014). This structure enables the examination of dual loadings, partitioning the variance of classroom quality dimensions between the general classroom quality factor and the domain-specific factors of Proactive Management and Routines and Cognitive Facilitation (Dunn & McCray, 2020). However, a primary limitation of the bifactor model is its more complex interpretation, particularly for dimensions that load onto both the general factor and a domain-specific factor, due to the partitioning of shared variance (Dunn & McCray, 2020). Nonetheless, the inclusion of a general factor may be essential for capturing a more comprehensive understanding of overall classroom quality in pre-K settings.

Multilevel Models

In addition to the CFAs, this analysis aimed to assess the impact of CLASS Pre-K factors on improvement in child outcomes from Fall to Spring. The comparison of the three-factor and bifactor models revealed no definitive superiority of either model. While the three-factor model demonstrated appropriate and intuitive factor loadings, the bifactor model showed slightly better fit statistics but included some problematic loadings. Given these findings, both models were used to examine the relationship between CLASS Pre-K factors and key outcomes on the DECA.

Intraclass Correlations

First, separate unconditional models with random intercepts were estimated for each of the two outcomes: change scores from Fall to Spring for DECA Total Protective Factors and Behavioral Concerns. The intraclass correlation coefficient (ICC) was calculated to determine the proportion of variance attributable to differences between classrooms versus within classrooms. For Total Protective Factors, 39% of the variance was attributed to differences between classrooms, while 61% was within classrooms. Similarly, for Behavioral Concerns, 33% of the variance was between classrooms, and 67% was within classrooms. These ICC values suggest that a substantial portion of the variance in child outcomes is influenced by classroom-level factors, underscoring the importance of accounting for classroom-level effects in the analysis.

Total Protective Factors Outcome

This analysis examined change scores for DECA Total Protective Factors from Fall to Spring using two sets of models. In the first set (Model 1), based on the CLASS Pre-K three-factor model, four separate models were assessed. Initially, each of the three factors—Emotional Support, Classroom Organization, and Instructional Support—was evaluated independently, followed by a fourth model that combined all three factors. In the second set (Model 2), derived from the CLASS Pre-K bifactor model, two models were assessed: one including only the general factor and the other incorporating both the general factor and the domain-specific factors. Across all models, the classroom average of Total Protective Factors was included as a covariate to control for baseline differences, and a set of dummy-coded variables to account for classrooms nested in schools. Classroom was specified as a random effect to capture clustering within classrooms (i.e., a random intercept). The factors of CLASS Pre-K served as the primary predictors of interest. Table 9 presents the correlation matrix for the CLASS Pre-K three-factor model, and the full results for each model are presented in Table 10.

Model 1: Evaluating the Predictive Capacity of the CLASS Pre-K Three-Factor Model for Changes in Total Protective Factors from Fall to Spring.

Model 1.1.

Model 1.1 included the predictor of Emotional Support. The average change in Total Protective Factors scores from Fall to Spring is estimated to be 4.80 points (SE = 2.20, $t = 2.18$, $p < .03$), when all predictors are held at 0, the dummy-coded variables that represent Head Start centers are at their reference group, and the Fall classroom average of Total Protective Factors is set at the sample mean. The predictor of Emotional Support is associated with a positive but non-significant increase in change scores for Total Protective Factors from Fall to Spring. A one-unit increase in Emotional Support is associated with a 0.50-point increase in the change scores (SE = 0.82, $t = .62$, $p = .54$). Model 1.1 explains 61% of the variance at the classroom level (i.e., of the between-classroom variance, 61% is accounted for by the predictors).

Model 1.2.

Model 1.2 included the predictor of Instructional Support. The average change in the Total Protective Factors scores from Fall to Spring is estimated to be 4.76 points (SE = 2.22, $t = 2.14$, $p = .03$), when all predictors are held at 0, the dummy-coded variables that represent Head Start centers are at their reference group, and the Fall classroom average of Total Protective Factors is set at the sample mean. The predictor of Instructional Support is associated with a positive but non-significant increase in the change scores for Total Protective Factors from Fall to Spring. A one-unit increase in Instructional Support is associated with a 0.30-point increase in the change scores (SE = 0.79, t = 0.38, p = .71). Model 1.2 explains 61% of the variance at the classroom level (i.e., of the between-classroom variance, 61% is accounted for by the predictors).

Model 1.3.

Model 1.3 included the predictor of Classroom Organization. The average change in the Total Protective Factors scores from Fall to Spring is estimated to be 4.79 points (SE 2.19, t = 2.18, p < .03), when all predictors are held at 0, the dummy-coded variables that represent Head Start centers are at their reference group, and the Fall classroom average of Total Protective Factors is set at the sample mean. The predictor of Classroom Organization is associated with a positive but non-significant increase in the change scores for Total Protective Factors from Fall to Spring. A one-unit increase in Classroom Organization is associated with a 0.59-point increase in the change scores (SE = 0.83, t = 0.71, p = .48). Model 1.3 explains 61% of the variance at the classroom level (i.e., of the between-classroom variance, 61% is accounted for by the predictors).

Model 1.4.

Model 1.4 included the predictors of Emotional Support, Instructional Support, and Classroom Organization that formulate the three-factor model of CLASS Pre-K. The average change in the Total Protective Factors scores from Fall to Spring for kids is estimated to be 4.96 points (SE = 2.24, t = 2.21, p < .03 when all predictors are held at 0, the dummy-coded variables that represent Head Start centers are at their reference group, and the Fall classroom average of Total Protective Factors is set at the sample mean. A one unit increase in the predictor of Emotional Support is associated with a 3.73-point non-significant decrease in the change scores for Total Protective Factors from Fall to Spring (SE = 6.07, t = -0.62, p = .54). A one unit increase in the predictor of Instructional Support is associated with a 0.38-point non-significant decrease in the change scores for Total Protective Factors from Fall to Spring (SE = 1.30, t =

-0.29, $p = .77$). Finally, a one unit increase in the predictor of Classroom Organization is associated with a 4.53-point non-significant increase in the change scores for Total Protective Factors from Fall to Spring (SE = 6.51, $t = 0.70$, $p = .49$). Model 1.4 explains 61% of the variance at the classroom level (i.e., of the between-classroom variance, 61% is accounted for by the predictors).

Model 2: Evaluating the Predictive Capacity of the CLASS Pre-K Bifactor Model for Changes in Total Protective Factors from Fall to Spring.

Model 2.1.

Model 2.1 included the predictor of Responsive Teaching, the general factor. The average change in the Total Protective Factors scores from Fall to Spring is estimated to be 4.82 points (SE = 2.20, $t = 2.19$, $p < .03$), when all predictors are held at 0, the dummy-coded variables that represent Head Start centers are at their reference group, and the Fall classroom average of Total Protective Factors is set at the sample mean. The predictor of Responsive Teaching is associated with a positive but non-significant increase in the change scores for Total Protective Factors from Fall to Spring. A one-unit increase in Responsive Teaching is associated with a 0.48-point increase in the change score (SE = 0.81, $t = 0.59$, $p = .55$). Model 2.1 explains 61% of the variance at the classroom level (i.e., of the between-classroom variance, 61% is accounted for by the predictors).

Model 2.2.

Model 2.2 included Responsive Teaching, Proactive Management and Routines, and Cognitive Facilitation as predictors, representing both the general and two domain-specific factors of the bifactor model of CLASS Pre-K. The average change in the Total Protective Factors scores from Fall to Spring is estimated to be 5.59 points (SE = 2.52, $t = 2.22$, $p < .03$), when all predictors are held at 0, the dummy-coded variables that represent Head Start centers are at their reference group, and the Fall classroom average of Total Protective Factors is set at the sample mean. A one-unit increase in the predictor of Responsive Teaching is associated with a 0.50-point non-significant increase in the change scores for Total Protective Factors from Fall to Spring (SE = 0.89, t = 0.56, p = .57). A one-unit increase in the predictor of Proactive Management and Routines is associated with a 0.91-point non-significant decrease in the change scores for Total Protective Factors (SE = 1.27, t = -0.71, p = .48). A one-unit increase in the predictor of Cognitive Facilitation is associated with a .05-point non-significant decrease in the change scores for Total Protective Factors (SE = 1.04, t = -0.05, p = .96). Model 2.2 explains 61% of the variance at the classroom level (i.e., of the between-classroom variance, 61% is accounted for by the predictors).

Behavioral Concerns Outcome

The remaining models included in this analysis examined change scores for DECA Behavioral Concerns from Fall to Spring using two sets of models. In the first set (Model 3), based on the CLASS Pre-K three-factor model, four separate models were assessed. Initially, each of the three factors—Emotional Support, Classroom Organization, and Instructional Support—was evaluated independently, followed by a fourth model that combined all three factors. In the second set (Model 4), derived from the CLASS Pre-K bifactor model, two models were assessed: one including only the general factor and the other incorporating both the general factor and the domain-specific factors. Across all models, the classroom average of Behavioral Concerns was included as a covariate to control for baseline differences, and a set of dummy-coded variables to account for classrooms nested in schools. Classroom was specified as a random effect to capture clustering within classrooms (i.e., a random intercept). The factors of

CLASS Pre-K served as the primary predictors of interest. The full results for each model are presented in Table 11.

Model 3: Evaluating the Predictive Capacity of the CLASS Pre-K Three-factor Model for Changes in Behavioral Concerns from Fall to Spring.

Model 3.1.

Model 3.1 included the predictor of Emotional Support. The average change in the Behavioral Concerns scores from Fall to Spring is estimated to be -1.30 points (SE = 1.97, t = -0.66, p = .51 when all predictors are held at 0, the dummy-coded variables that represent Head Start centers are at their reference group, and the Fall classroom average of Behavioral Concerns is set at the sample mean. The predictor of Emotional Support is associated with a negative but non-significant increase in the change score for Behavioral Concerns from Fall to Spring. A one-unit increase in Emotional Support is associated with a 0.03-point decrease in the change scores (SE = 0.87, t = -0.03, p = .98). Model 3.1 explains 54% of the variance at the classroom level (i.e., of the between-classroom variance, 54% is accounted for by the predictors).

Model 3.2.

Model 3.2 included the predictor of Instructional Support. The average change in the Behavioral Concerns scores from Fall to Spring is estimated to be -1.07 points (SE = 1.99, t = -0.54, p = .59), when all predictors are held at 0, the dummy-coded variables that represent Head Start centers are at their reference group, and the Fall classroom average of Behavioral Concerns is set at the sample mean. The predictor of Instructional Support is associated with a non-significant decrease in the change scores for Behavioral Concerns from Fall to Spring. A one-unit increase in Instructional Support is associated with a 0.48-point non-significant decrease in the change scores (SE = 0.75, t = -0.64, p = .52). Model 3.2 explains 54% of the variance at the classroom level (i.e., of the between-classroom variance, 54% is accounted for by the predictors).

Model 3.3.

Model 3.3 included the predictor of Classroom Organization. The average change in the Behavioral Concerns scores from Fall to Spring is estimated to be -1.28 points (SE = 1.97, t = -0.65, p = .52), when all predictors are held at 0, the dummy-coded variables that represent Head Start centers are at their reference group, and the Fall classroom average of Behavioral Concerns is set at the sample mean. The predictor of Classroom Organization is associated with a non-significant decrease in the change scores for Behavioral Concerns from Fall to Spring. A one-unit increase in Classroom Organization is associated with a 0.14-point non-significant decrease in the change scores (SE = 0.86, t = -0.16, p = .87). Model 3.3 explains 54% of the variance at the classroom level (i.e., of the between-classroom variance, 54% is accounted for by the predictors).

Model 3.4.

Model 3.4 included the predictors of Emotional Support, Instructional Support, and Classroom Organization that formulate the three-factor model of CLASS Pre-K. The average change in the Behavioral Concerns scores from Fall to Spring is estimated to be -1.12 points (SE = 1.99, t = -0.56, p = .58), when all predictors are held at 0, the dummy-coded variables that represent Head Start centers are at their reference group, and the Fall classroom average of Behavioral Concerns is set at the sample mean. A one unit increase in the predictor of Emotional Support is associated with a 3.90-point non-significant increase in the change scores for Behavioral Concerns from Fall to Spring (SE = 7.04, t = 0.55, p = .58). A one unit increase in the predictor of Instructional Support is associated with a 0.63-point non-significant decrease in the change scores for Behavioral Concerns from Fall to Spring (SE = 1.14, t = -0.55, p = .58). Finally, a one unit increase in the predictor of Classroom Organization is associated with a 3.45-point non-significant decrease in the change scores for Behavioral Concerns from Fall to Spring (SE = 7.33, t = -0.47, p = .64). Model 3.4 explains 55% of the variance at the classroom level (i.e., of the between-classroom variance, 55% is accounted for by the predictors), which is a 1% increase from the models including a single factor.

Model 4: Evaluating the Predictive Capacity of the CLASS Pre-K Bifactor Model for Changes in Behavioral Concerns from Fall to Spring.

Model 4.1.

Model 4.1 incorporated Responsive Teaching as a predictor, representing the general factor from the bifactor model of CLASS Pre-K. The average change in the Behavioral Concerns scores from Fall to Spring is estimated to be -1.29 points (SE = 1.97, t = -0.66, p = .51), when all predictors are held at 0, the dummy-coded variables that represent Head Start centers are at their reference group, and the Fall classroom average of Behavioral Concerns is set at the sample mean. The predictor of Responsive Teaching is associated with a non-significant decrease in the change score for Behavioral Concerns from Fall to Spring. A one-unit increase in Responsive Teaching is associated with a 0.07-point non-significant decrease in the change scores (SE = 0.84, t = -0.08, p = .93). Model 4.1 explains 54% of the variance at the classroom level (i.e., of the between-classroom variance, 54% is accounted for by the predictors).

Model 4.2.

Model 4.2 included Responsive Teaching, Proactive Management and Routines, and Cognitive Facilitation as predictors, representing both the general and two domain-specific factors of the bifactor model of CLASS Pre-K. The average change in the Behavioral Concerns scores from Fall to Spring is estimated to be -1.67 points (SE = 2.30, t = -0.73, p = .47), when all predictors are held at 0, the dummy-coded variables that represent Head Start centers are at their reference group, and the Fall classroom average of Behavioral Concerns is set at the sample mean. A one-unit increase in the predictor of Responsive Teaching is associated with a 0.17-point non-significant increase in the change scores for Behavioral Concerns from Fall to Spring (SE = 0.89, t = 0.19, p = .85). A one-unit increase in the predictor of Proactive Management and Routines is associated with a 0.79-point non-significant increase in the change scores for Behavioral Concerns (SE = 1.42, t = 0.56, p = .58). A one-unit increase in the predictor of Cognitive Facilitation is associated with a 0.74-point non-significant decrease in the change scores for Behavioral Concerns (SE = 0.90, t = -0.82, p = .41). Model 4.2 explains 56% of the variance at the classroom level (i.e., of the between-classroom variance, 56% is accounted for by the predictors), which is a 2% increase from the model only including the general factor.

Summary of MLMs

The multilevel models in this dissertation yielded comparable results across all predictors for both outcomes, with non-significant effects and small effect sizes. Specifically, no evidence was found that the CLASS Pre-K factor scores, whether from the three-factor or bifactor model, can significantly predict changes from Fall to Spring in DECA Behavioral Concerns or Total Protective Factors scores for Head Start children. Additionally, each model revealed significant variation in change scores across classrooms beyond the fixed predictors, as well as substantial individual-level variability.

Sensitivity Analysis

Due to the high correlation between Emotional Support and Classroom Organization in the three-factor model of CLASS Pre-K, a two-factor model was tested as a sensitivity analysis.

This two-factor model combined Emotional Support and Classroom Organization into a single factor and retained Instructional Support as a separate factor. Results indicated that the two-factor model demonstrated poor fit to the data (CFI = .92, TLI = .89, RMSEA = .13, SRMR = .08). Additionally, the dimensions of *Positive Climate* and *Negative Climate*, now loading on the combined Emotional Support/Classroom Organization factor, exhibited weak factor loadings (λ = .44 and λ = –.09, respectively).

Multilevel models (MLMs) were also conducted to examine whether the two-factor structure predicted changes in DECA scores from Fall to Spring. The combined Emotional Support/Classroom Organization factor did not significantly predict changes in Total Protective Factors (*b* = 0.53, SE = 1.27, t = 0.41, p = .68) or Behavioral Concerns (*b* = 0.41, SE = 1.08, t = 0.38, p = .71). Similarly, Instructional Support was not a significant predictor of change in Total Protective Factors (*b* = –0.23, SE = 1.19, t = –0.19, p = .85) or Behavioral Concerns (*b* = –0.69, SE = 1.02, t = –0.68, p = .50) when included as the second factor in the two-factor model of CLASS Pre-K. These results indicate that the two-factor model failed to address the issues of poor model fit and limited predictive validity observed in the original structure.

DISCUSSION

Overview of CLASS Pre-K Factor Analyses

This dissertation first investigated the factor structure of CLASS Pre-K using confirmatory factor analyses (CFAs) to compare the traditional three-factor model (Pianta et al., 2008b) with an alternative bifactor model (Hamre et al., 2014). The three-factor model conceptualizes Emotional Support, Classroom Organization, and Instructional Support as distinct but interrelated domains of teacher–child interaction (Pianta et al., 2008b). The bifactor model was developed to address empirical challenges in the three-factor structure by introducing a general factor, Responsive Teaching, alongside two domain-specific factors: Proactive Management and Routines, and Cognitive Facilitation (Hamre et al., 2014). Although both models are theoretically grounded in developmental frameworks, the findings from this dissertation reveal substantial structural and conceptual limitations in both models when applied to Head Start classrooms serving culturally and linguistically diverse populations (Gordon & Peng, 2020; Ng et al., 2021).

Model Fit

Three-Factor Model.

Before interpreting model fit, it is crucial to acknowledge the extremely high correlation between Emotional Support and Classroom Organization in the three-factor model of CLASS Pre-K (r = .96), which undermines the assumption that these constructs are empirically distinct (Brown, 2015; Hamre et al., 2014; Sandilos et al., 2016). Such strong collinearity suggests that the two factors may be measuring the same underlying construct, threatening the model's discriminant validity—the extent to which theoretically separate domains are actually distinct

(Brown, 2015; Campbell & Fiske, 1959). This overlap complicates the interpretation of individual factor scores and raises concerns about their utility in applied settings such as professional development or classroom coaching. To explore this issue, a sensitivity analysis was conducted using a two-factor model that combined Emotional Support and Classroom Organization into a single factor (Sabol et al., 2013), but this approach did not yield improved model fit or predictive utility. Fit indices for the original three-factor model remained suboptimal ($\chi^2(32, n = 86) = 78.07$, CFI = .92, TLI = .89, RMSEA = .13, SRMR = .08), falling short of conventional thresholds for good model fit (CFI and TLI ≥ .95, RMSEA < .06, SRMR < .08; Bentler, 1990; Hu & Bentler, 1999). In particular, the elevated RMSEA value suggests poor structural fit and indicates that the model may not adequately capture the distinct dimensions of classroom quality it purports to measure.

Despite the poor fit, all dimensions in the three-factor model demonstrated acceptable standardized factor loadings ($\lambda > .40$), which indicates that each observed variable was moderately to strongly associated with its respective latent factor (Brown. 2015; Kline, 2016). Notably, the *Negative Climate* dimension, which has frequently shown unstable or weak loadings in prior research, loaded at $\lambda = -.53$, meeting the statistical threshold but raising questions about interpretability due to the inconsistent behavior across samples (Leyva et al., 2015; Pakarinen et al., 2010). A factor loading reflects how well an observed indicator represents its latent construct, and values below .40 are typically considered weak (Brown, 2015). Thus, although the three-factor model's dimensions met this criterion for factor loadings, the overall structural alignment remained poor.

In addition to the concerningly high correlation between Emotional Support and Classroom Organization, Instructional Support consistently emerged as having the most conceptual weaknesses, both in this study and in the broader literature. For example, Instructional Support tends to display lower classroom average scores, weaker factor loadings, and restricted variance compared to the other domains (Gordon & Peng, 2020; Leyva et al., 2015). This pattern is frequently observed in early childhood settings, particularly in programs serving children from low-income or multilingual backgrounds (Cadima et al., 2018; Downer et al., 2012). The domain of Instructional Support emphasizes advanced linguistic and cognitive strategies, such as scaffolding, inferencing, and extended feedback loops (Pianta et al., 2008b), which may be more challenging to observe consistently in pre-K classrooms, especially when teachers are responding to children' basic developmental, behavioral, or socio-emotional needs (Hu et al., 2016; Li et al., 2020).

Research indicates that Instructional Support scores on CLASS Pre-K tend to cluster at the low end of the scale, especially in under-resourced pre-K programs such as Head Start (Burchinal et al., 2010; Cadima et al., 2018; Gordon & Peng, 2020; Hu et al., 2016; Leyva et al., 2015). Many classrooms score only in the low to mid-range on this domain, resulting in a skewed distribution with restricted variance (Burchinal et al., 2010). This lack of variability complicates measurement, as the tool has limited sensitivity to detect fine-grained differences between classrooms or growth over time (Cadima et al., 2018; Gordon & Peng, 2020). Furthermore, this pattern raises concerns about content validity, which refers to how well an instrument captures the full scope of the construct it intends to measure (Brown, 2015). If Instructional Support, as currently defined, is not consistently observable in diverse pre-K classrooms, this may reflect a mismatch between what is measured and what is both feasible and developmentally appropriate in these settings (Hu et al., 2016; Leyva et al., 2015). For example, certain forms of abstract or scaffolded instruction emphasized by the CLASS framework may be rare in classrooms serving low-income or multilingual children—not necessarily due to poor teaching, but due to differences in developmental needs, cultural values, and instructional priorities (Hu et al., 2016; Leyva et al., 2015).

Cultural context also plays a significant role in shaping how teachers engage in instructional exchanges as measured by the factor of Instructional Support (Downer et al., 2012). As such, the CLASS framework may not fully capture meaningful instruction in multilingual or culturally diverse classrooms (Downer et al., 2012; Ng et al., 2021). Research has shown that culturally grounded teaching practices—such as use of home language, storytelling traditions, or subtle forms of cognitive support—may not align with CLASS Pre-K's operational definitions of high-quality instruction (Downer et al., 2012; Ng et al., 2021; Hu et al., 2016). These differences may lead to underestimation of instructional quality in classrooms that do not match the cultural norms on which the tool was developed (Downer et al., 2012; Ng et al., 2021). For instance, teachers working in multilingual or culturally diverse classrooms may engage in high-quality instructional exchanges that are less explicit or structured than those emphasized by CLASS Pre-K, yet still meaningful and effective for children's learning (Leyva et al., 2015). Together, these findings suggest the need for caution in interpreting low Instructional Support scores and support the call for more culturally responsive or contextually aligned measurement tools in Head Start and similar early childhood education settings.

Bifactor Model.

The bifactor model, proposed as an alternative to address high inter-factor correlations, did not resolve the model fit issues that appeared with the three-factor model in this dissertation (Gordon & Peng, 2020; Hamre et al., 2014). In the bifactor model, a general factor of Responsive Teaching is posited to account for shared variance across all CLASS Pre-K dimensions, while domain-specific factors (e.g., Proactive Management and Routines, Cognitive Facilitation) account for unique aspects of classroom interaction (Hamre et al., 2014). However, several items in this model failed to meet the standard loading threshold ($\lambda < .40$), indicating that these observed variables were weakly related to their designated latent constructs (Brown, 2015; Gordon & Peng, 2020; Ng et al., 2021). Additionally, estimation issues such as a negative factor loading for *Behavior Management* on Proactive Management and Routines suggest an unstable and potentially misspecified model (Li et al., 2020). These types of estimation problems are often symptomatic of conceptual issues in the model, such as poorly defined latent variables or overlapping item content (Bihler et al., 2018; Karlsen et al., 2023).

Previous studies investigating the bifactor model of CLASS Pre-K have also encountered estimation issues. For example, Hamre et al. (2014) initially proposed a bifactor model with three domain-specific factors but encountered convergence problems and revised the model to include only two specific factors. Even in this revised structure, several items did not load significantly on their intended domain-specific factors, and associations with child outcomes remained small (Hamre et al., 2014). Similarly, Gordon and Peng (2020), using large, nationally representative Head Start samples, reported poor model fit and estimation issues with the bifactor model structure, concluding that neither the standard three-factor model nor the bifactor model provided a strong empirical foundation. Ng et al. (2021) found that a bifactor model with one general and two specific factors fit better than the three-factor model in Singaporean preschool and early intervention classrooms but still exhibited partial metric invariance and cross-loading issues, raising concerns about generalizability and construct clarity. Collectively, these studies suggest that while the bifactor model offers a conceptual solution to multicollinearity, it often fails to provide a stable and interpretable measurement structure in practice (Gordon & Peng, 2020; Hamre et al., 2014; Ng et al., 2021). As such, the empirical challenges encountered in this dissertation reinforce broader concerns about the validity and utility for the bifactor model of the CLASS Pre-K in diverse educational contexts (Bihler et al., 2018; Karlsen et al., 2023).

Model Adjustments Found in Prior Research.

Many studies addressing poor initial model fit have introduced post hoc modifications to their confirmatory factor analyses of CLASS Pre-K data, such as correlating residuals between dimensions or allowing indicators to load onto multiple latent factors (Hamre et al., 2014; Karlsen et al., 2023; Ng et al., 2021). For example, Karlsen et al. (2023) improved model fit by correlating residuals between *Negative Climate* and *Behavior Management* in a Norwegian sample. Similarly, Ng et al. (2021) reported using cross-loadings and residual correlations in bifactor models to better account for observed interaction patterns. Hamre et al. (2014) also proposed a bifactor structure incorporating a general Responsive Teaching factor, which permitted each CLASS dimension to load onto both general and specific factors, resulting in improved model fit compared to the original three-domain framework.

Although these modifications often lead to more favorable statistical indices, they are typically atheoretical and diverge from the original CLASS Pre-K conceptual framework, which was designed around ten distinct dimensions grouped into three theoretically defined domains: Emotional Support, Classroom Organization, and Instructional Support (Pianta et al., 2008b). Each indicator was intended to load onto a single, conceptually distinct domain. When modifications such as cross-loadings or correlated residuals are introduced without a guiding theory, the model begins to deviate from its foundational structure, raising questions about whether the revised model is still measuring what it was originally designed to assess.

This concern reflects a broader issue related to validity. As Kline (2016) explains, improved statistical fit does not ensure that the underlying constructs are conceptually meaningful. Model fit indices indicate only how well the proposed structure reproduces the observed data, not whether the theoretical constructs are valid representations of the phenomena in question (Kline, 2016). When modifications are driven solely by statistical outputs rather than theoretical rationale, there is a risk that the model will reflect sample-specific idiosyncrasies rather than generalizable constructs (Kline, 2016). In the context of CLASS Pre-K, this could mean that factor scores no longer meaningfully capture teacher–child interaction quality as originally defined.

Such issues point directly to concerns about construct validity, which refers to the degree to which an instrument accurately measures the theoretical constructs it purports to assess (Campbell & Fiske, 1959). Establishing construct validity requires demonstrating both convergent and discriminant evidence, and post hoc modifications that blur the boundaries between latent constructs may compromise this balance (Campbell & Fiske, 1959). For instance, cross-loadings or correlated errors can artificially inflate associations between domains, weakening evidence that each domain reflects a unique aspect of classroom interaction.

These concerns are further reinforced by empirical findings. Hamre et al. (2014) reported that the three CLASS Pre-K factors are often highly correlated, which raises doubts about their empirical distinctiveness. Introducing additional cross-loadings or residual covariances without a clear theoretical rationale may only exacerbate this overlap, further entangling the constructs (Hamre et al., 2014). Kline (2016) cautions that such statistically driven adjustments may produce factors that are difficult to interpret and no longer align with the instrument's conceptual goals.

This lack of clarity has practical consequences. Downer et al. (2012) emphasize that in applied contexts—such as professional development, teacher coaching, or accountability systems—it is essential that each CLASS Pre-K factor score reliably reflects a distinct and interpretable construct. When model structures vary across studies or depend on extensive sample-specific modifications, the consistency and usefulness of those scores are undermined (Downer et al., 2012; Hamre et al., 2014). As a result, educators and program leaders may lose confidence in the diagnostic value of domain scores, particularly when it is unclear whether low scores reflect true instructional weaknesses or measurement artifacts (Downer et al., 2012; Hamre et al., 2014).

Beyond interpretability, the use of extensive post hoc modifications raises concerns about model parsimony and generalizability. As Kline (2016) notes, each additional parameter—whether a residual correlation or a cross-loading—increases model complexity and the risk of overfitting to a particular sample. Downer et al. (2012) similarly argue that modifications aimed at improving model fit in one context may fail to generalize across diverse classroom settings. This risk is evident in Karlsen et al. (2023), who found that the modifications needed to improve model fit in their Norwegian sample might not be necessary—or even appropriate—in other cultural contexts. Ng et al. (2021) also caution that such modifications may not transfer reliably across samples with different linguistic or instructional characteristics.

Although model adjustments can produce more acceptable fit statistics, they highlight deeper theoretical limitations in the measurement model. Both Hamre et al. (2014) and Karlsen et al. (2023) argue that the need for substantial post hoc modifications suggests that the standard CLASS framework may not fully capture the complexity or variability of teacher–child interactions across diverse educational contexts. In such cases, pursuing statistical fit without attention to theoretical coherence may undermine the conceptual integrity and practical utility of the CLASS Pre-K tool (Hamre et al., 2014; Karlsen et al., 2023; Kline, 2016).

Conceptual and Structural Challenges of the CLASS Framework

Theoretical Assumptions.

The shortcomings identified in this dissertation, guided by findings from previous factor analytic studies, are not merely statistical artifacts; rather, they reveal deeper conceptual tensions within the CLASS framework (Gordon & Peng, 2020; Ng et al., 2021). At the heart of the CLASS framework is the assumption that teacher–child interactions can be categorized into three distinct and equally important domains: Emotional Support, Classroom Organization, and Instructional Support (Pianta et al., 2008b; Hamre et al., 2014). These domains are theorized to capture stable, trait-like aspects of teacher behavior that are consistent across time and context (Hamre et al., 2014; Ng et al., 2021).

However, this assumption has been increasingly questioned by evidence suggesting that observed teacher behaviors may vary significantly depending on the situation. For example, a teacher might display high emotional sensitivity during a morning meeting but shift toward behavior management during transitions or challenging instructional moments (Hamre et al., 2014). This situational variability calls into question whether CLASS Pre-K scores reflect enduring teaching traits or moment-to-moment adaptations to dynamic classroom conditions (Gordon & Peng, 2020; Pakarinen et al., 2010). Viewed through this lens, CLASS Pre-K may be better understood as a measure of context-sensitive behavior rather than stable latent constructs (Karlsen et al., 2023; Ng et al., 2021).

Empirical Inconsistencies.

Despite its widespread use, the alignment between CLASS Pre-K's theoretical framework and empirical evidence has proven inconsistent, prompting ongoing scrutiny of its factor structure. A persistent empirical challenge in CLASS Pre-K research is the high correlation between Emotional Support and Classroom Organization, which has been documented in multiple studies and replicated in this dissertation, often exceeding $r = .90$ (Li et al., 2020; Pakarinen et al., 2010). Such strong associations raise concerns about the discriminant validity of these factors and suggest that they may not function as distinct constructs in practice (Bihler et al., 2018; Gordon & Peng, 2020). Some researchers argued that these factors together may reflect a broader construct such as emotional-climatic quality, which integrates affective warmth, responsiveness, and behavioral structure (Cadima et al., 2018; Hamre et al., 2014). Indeed, emotionally attuned teachers are often more effective at managing classrooms, as both require responsiveness to children's signals and needs (Downer et al., 2012; Gordon & Peng, 2020).

In contrast, Instructional Support typically emerges as the most distinct yet underdeveloped domain in early childhood settings (Gordon & Peng, 2020; Leyva et al., 2015). Empirical findings consistently show that this domain has lower factor loadings, reduced variability, and scores clustered toward the low end of the scale (Cadima et al., 2018; Li et al., 2020). These results may reflect the high cognitive demands of Instructional Support behaviors—such as scaffolding, inferential questioning, and metacognitive dialogue—which are often deprioritized in classrooms where socio-emotional needs take precedence (Downer et al., 2012; Hu et al., 2016). Moreover, the instructional strategies emphasized by CLASS Pre-K may not align with all culturally responsive or developmentally appropriate pedagogies (Leyva et al., 2015; Ng et al., 2021).

Implications for Composite Scoring and Practical Use.

The structural and conceptual inconsistencies outlined above have important implications for how CLASS Pre-K scores are used in practice. One particularly problematic area is the use of composite scores, which combine ratings (typically by averaging) from the three-factor model factor scores into a single index of classroom quality (Hamre et al., 2014; Karlsen et al., 2023). This approach assumes that the factors are conceptually independent and empirically balanced in their contributions to classroom quality. However, if the factors are highly correlated or perform unevenly across contexts, composite scoring may obscure meaningful differences and lead to misleading interpretations (Cadima et al., 2018; Downer et al., 2012; Li et al., 2020).

The possibility for misleading interpretations of CLASS Pre-K scores is especially concerning in high-stakes contexts, where they are used to inform teacher evaluations, program funding, and accountability decisions (Gordon & Peng, 2020; Hamre et al., 2014). Aggregating across conceptually or culturally misaligned domains risks penalizing educators or programs that may excel in contextually appropriate ways not fully captured by the CLASS framework (Ng et al., 2021; Leyva et al., 2015). As such, both the scientific integrity and practical utility of the tool are compromised when its scores are not grounded in a valid and coherent measurement model (Cadima et al., 2018; Karlsen et al., 2023).

Cultural Considerations

Sociocultural Misalignment and Measurement Validity.

A critical issue underlying these structural problems is the question of sociocultural relevance—specifically, whether the CLASS Pre-K constructs are culturally universal or context-dependent (Downer et al., 2012; Gordon & Peng, 2020; Hu et al., 2016). For example, the expression of *Teacher Sensitivity*, a dimension of Emotional Support, may differ across cultural groups, leading to variation in how these behaviors are enacted and interpreted (Bihler et al., 2018; Leyva et al., 2015). Similarly, expectations for Classroom Organization, such as norms around structure, autonomy, or group responsibility, are shaped by cultural context and may not align with CLASS Pre-K dimensions (Cadima et al., 2018; Hu et al., 2016). When the behavioral markers of CLASS Pre-K factors are culturally variable, their validity as a cross-cultural tool is compromised (Gordon & Peng, 2020; Ng et al., 2021).

International studies reinforce these concerns regarding cultural variability, as researchers in Germany, Portugal, Finland, Chile, and China have repeatedly struggled to fit the original three-domain structure to their data (Bihler et al., 2018; Cadima et al., 2018; Hu et al., 2016; Leyva et al., 2015; Pakarinen et al., 2010). In many cases, dimensions such as *Positive Climate, Teacher Sensitivity*, and *Negative Climate* required modification or exclusion, indicating inconsistent conceptual and empirical utility across cultural contexts (Bihler et al., 2018; Leyva et al., 2015). These challenges suggest the CLASS framework lacks structural invariance and should not be assumed to apply universally without thoughtful adaptation (Downer et al., 2012; Ng et al., 2021).

Implications for Practice and Equity.

The sociocultural and structural limitations of CLASS Pre-K have meaningful implications in applied settings, particularly within programs like Head Start that serve racially, culturally, and linguistically diverse populations (Downer et al., 2012; Hu et al., 2016). In these classrooms, reliable and culturally responsive measurement tools are essential for ensuring equitable assessments and supports (Cadima et al., 2018; Gordon & Peng, 2020). However, when

Emotional Support and Classroom Organization show high overlap, they may fail to provide differentiated feedback to teachers, limiting the tool's utility in professional development and coaching (Karlsen et al., 2023; Gordon & Peng, 2020).

Misalignment between CLASS Pre-K dimensions and cultural norms may lead to biased assessments, particularly when expectations around warmth, enthusiasm, or behavioral control do not reflect the norms of the communities being assessed (Leyva et al., 2015; Ng et al., 2021). These biases can result in inaccurate evaluations, misdirected coaching, and inequitable resource distribution—outcomes that are especially problematic in programs committed to promoting educational equity (Downer et al., 2012; Hu et al., 2016).

Evidence from this dissertation, which drew from a Head Start sample in which 66% of children were racially minoritized and 21% were dual language learners, further emphasizes the need for culturally valid tools (Downer et al., 2012; Hu et al., 2016; Leyva et al., 2015). Despite CLASS Pre-K's widespread use in federal accountability systems, few studies have systematically tested its structural validity in classrooms where diversity is the norm rather than the exception (Gordon & Peng, 2020). These findings suggest that current scoring and domain structures may fail to capture the relational and instructional practices most salient in these settings (Cadima et al., 2018; Ng et al., 2021).

Recommendations for CLASS Pre-K Refinement and Cultural Responsiveness

These limitations of CLASS Pre-K indicated in this dissertation—evidenced in poor model fit, overlapping factors, and mismatches with diverse classroom practices—have generated strong calls for refinement (Bihler et al., 2018; Gordon & Peng, 2020; Karlsen et al., 2023; Pakarinen et al., 2010). To ensure that CLASS Pre-K functions as a valid and equitable measure of teaching quality across varied settings, meaningful revisions must be made to its factor structure, scoring practices, and cultural adaptability (Downer et al., 2012; Hu et al., 2016; Leyva et al., 2015).

Refining the Conceptual Framework and Measurement Model.

Given persistent empirical and conceptual challenges, there is growing consensus that the CLASS Pre-K framework requires significant revision to fulfill its intended purpose of producing equitable, valid, and actionable assessments (Gordon & Peng, 2020; Hamre et al., 2014; Ng et al., 2021). In particular, the frequent overlap between domains such as Emotional Support and Classroom Organization undermines the assumption of domain independence and calls into question the validity of composite scoring (Cadima et al., 2018; Karlsen et al., 2023). Future revisions should include efforts to re-clarify domain definitions and test alternative structural models that reflect the dynamic, interrelated nature of classroom interactions (Gordon & Peng, 2020; Pakarinen et al., 2010).

Additionally, researchers and practitioners should explore new modeling approaches that account for situational variability in teacher behavior and better capture how instructional quality unfolds across different classroom contexts (Downer et al., 2012; Hu et al., 2016). These efforts would not only strengthen the psychometric soundness of the tool but also ensure that CLASS Pre-K more accurately reflects the fluid realities of early childhood classrooms (Ng et al., 2021; Pianta et al., 2008b).

Centering Cultural Responsiveness in Future Revisions.

Beyond measurement refinements, there is an urgent need to ensure that CLASS Pre-K is responsive to the cultural and linguistic diversity of contemporary classrooms. Continuing to apply a standardized observational framework across culturally varied settings risks erasing community-rooted teaching strengths and promoting deficit-oriented evaluations (Cadima et al., 2018; Gordon & Peng, 2020). Classrooms serving dual language learners, racially minoritized children, and children from historically marginalized communities require tools that reflect their values, pedagogical norms, and interactional patterns (Downer et al., 2012; Hu et al., 2016). Evidence from international and U.S.-based studies consistently shows that CLASS Pre-K dimensions—such as *Positive Climate* and *Teacher Sensitivity*—are interpreted and enacted differently across cultures, potentially distorting ratings when used in non-dominant contexts (Bihler et al., 2018; Hu et al., 2016; Leyva et al., 2015). To address this, future research should prioritize validating domain structures across diverse subgroups, including Head Start populations, multilingual learners, and classrooms embedded in distinct cultural traditions (Gordon & Peng, 2020; Hamre et al., 2014).

Expanding Indicators of Quality Through Cultural Adaptation.

Incorporating culturally grounded indicators of quality is a critical step toward building a more inclusive and accurate version of CLASS Pre-K (Cadima et al., 2018; Downer et al., 2012; Gordon & Peng, 2020; Hu et al., 2016; Leyva et al., 2015; Ng et al., 2021). Practices such as storytelling, the use of home language, relational approaches to authority, and culturally relevant routines are often central to instructional quality in many communities but may fall outside the scope of the current CLASS Pre-K factors (Cadima et al., 2018; Hu et al., 2016). Without these adaptations, CLASS Pre-K risks misrepresenting high-quality teaching and reinforcing narrow definitions of excellence rooted in dominant cultural norms (Gordon & Peng, 2020; Leyva et al., 2015).

Developing supplemental tools that align with localized interaction patterns may further improve fairness and interpretability (Downer et al., 2012; Hamre et al., 2014). These could include measures of bilingual scaffolding, culturally responsive pedagogy, or family engagement practices highly valued by communities but underrepresented in traditional CLASS Pre-K scoring (Gordon & Peng, 2020; Ng et al., 2021). Such tools could complement existing CLASS Pre-K dimensions, providing a completer and more representative picture of teaching quality in classrooms that serve culturally and linguistically diverse learners (Hu et al., 2016; Leyva et al., 2015). When observational tools better reflect the lived experiences and practices of children and teachers, the feedback they generate becomes more useful for coaching, professional development, and continuous quality improvement (Cadima et al., 2018; Karlsen et al., 2023). This alignment is especially crucial in high-stakes environments where CLASS Pre-K scores are used for program ratings and funding decisions (Downer et al., 2012; Hamre et al., 2014).

Recommendations for Research and Framework Development

The inconsistent or poor model fit observed across multiple studies points to conceptual limitations in how the CLASS Pre-K framework theorizes and operationalizes instructional quality (Hamre et al., 2014; Ng et al., 2021). Researchers have found that the standard three-domain structure—Emotional Support, Classroom Organization, and Instructional Support—is not consistently supported by empirical data (Hamre et al., 2014; Ng et al., 2021). Studies often report high inter-correlations among domains and better fit for alternative structures, such as bifactor or two-domain models (Hamre et al., 2014; Ng et al., 2021). For example, in large U.S. pre-K samples, bifactor models that included a general Responsive Teaching factor alongside specific domain factors showed improved model fit, raising questions about whether the original structure meaningfully captures distinct instructional processes (Hamre et al., 2014).

Similar concerns have been documented in large-scale Head Start data (Gordon & Peng, 2020). Gordon and Peng (2020) found that in nationally representative samples, alternative factor structures fit the data as well or better than the standard CLASS Pre-K configuration (Gordon &

Peng, 2020). Moreover, associations between CLASS Pre-K factor scores and children's academic or social-emotional outcomes were weak or nonsignificant, suggesting that the tool may not fully capture the aspects of teaching that matter most for child development in these settings (Gordon & Peng, 2020). These findings point to the urgent need for sustained theoretical refinement and empirical revalidation of the CLASS framework, particularly given its widespread use in diverse and under-resourced classrooms (Downer et al., 2012; Gordon & Peng, 2020).

Without this foundational work, the tool's ability to equitably assess and support early learning environments will remain in doubt (Cadima et al., 2018; Karlsen et al., 2023). In Portugal, for example, a two-domain model fit better than the standard three-factor structure, while in Chile and Finland, residual correlations had to be added just to achieve adequate fit (Cadima et al., 2018). Similarly, Karlsen et al. (2023) found that post hoc model modifications were necessary to obtain an acceptable fit in Norwegian classrooms, underscoring how cultural and contextual variation can challenge the applicability of a fixed measurement model (Karlsen et al., 2023). These examples highlight the risk of applying CLASS uniformly across settings without careful validation (Cadima et al., 2018; Karlsen et al., 2023).

Rather than relying solely on confirmatory factor analyses, future research should adopt exploratory, theory-building approaches that reflect the fluid and culturally embedded nature of classroom life (Karlsen et al., 2023; Ng et al., 2021). This includes the use of exploratory factor analyses, open-ended observation tools, and mixed-methods designs that allow new constructs to emerge organically (Ng et al., 2021). Participatory approaches that involve educators, families, and community members in the development and refinement of measurement tools can help ensure that the instruments reflect what matters most in local classroom contexts (Cadima et al., 2018; Hu et al., 2016).

Subgroup-specific validation studies also play a critical role in refining CLASS Pre-K for equitable use (Cadima et al., 2018; Hu et al., 2016). Hu et al. (2016), for instance, found that a bifactor model provided adequate fit in Chinese preschool classrooms but noted differences in the way some CLASS dimensions functioned (Hu et al., 2016). These findings suggest that measurement structures may need to be adapted to local teaching norms, interaction styles, or cultural expectations (Hu et al., 2016). Without such adaptations, tools like CLASS Pre-K risk misrepresenting or undervaluing the quality of teacher–child interactions in classrooms that differ from those used in the tool's original development (Downer et al., 2012; Gordon & Peng, 2020).

Together, these strategies—exploratory model building, participatory design, and context-specific validation—are essential for ensuring that CLASS Pre-K meaningfully and equitably captures instructional quality (Cadima et al., 2018; Karlsen et al., 2023; Ng et al., 2021). As it stands, the CLASS framework may still reflect the priorities and assumptions of its original development context, and its widespread application to settings such as Head Start requires a commitment to continuous theoretical and empirical refinement (Downer et al., 2012; Gordon & Peng, 2020). Only then can CLASS Pre-K be relied upon to inform equitable assessment, professional development, and policy across the diverse spectrum of early childhood education environments (Hamre et al., 2014; Gordon & Peng, 2020; Ng et al., 2021).

Multilevel Models

After examining the factor structure of CLASS Pre-K, this dissertation investigated the extent to which CLASS Pre-K factor scores, derived from both the three-factor and bifactor models, predicted gains in DECA Total Protective Factors and reductions in Behavioral Concerns from Fall to Spring in Head Start classrooms. Contrary to expectations, no significant associations were found between CLASS Pre-K factor scores and DECA change scores. These null findings may be partially attributed to limited statistical power, as the analytic sample included a small number of classrooms with available outcome data (Hedges & Rhoads, 2010).

ICCs

In addition to model estimates, since most research completed in the education field involves clustered designs (e.g., children nested within classrooms or schools), intraclass correlations (ICCs) must be accounted for, as they affect statistical power by increasing the similarity of outcomes within clusters (Hedges & Rhoads, 2010). The ICCs in this dissertation (ICC = .26 for Total Protective Factors and ICC = 23 for Behavioral Concerns) suggest that while classrooms play an important role in shaping child outcomes, individual differences among children account for a larger share of the variation in change scores from Fall to Spring (approximately 74% for Total Protective Factors and 77% for Behavioral Concerns). The higher proportion of within-classroom variance suggests that child-specific factors, such as prior skills, temperament, or home environment, may be more influential than classroom-wide characteristics. However, the remaining between-classroom variance underscores the continued importance of classroom quality in supporting child development. These findings highlight the need for a balanced approach to improving child outcomes in Head Start: while enhancing overall classroom quality remains essential, targeted interventions for at-risk children may have an even greater impact in addressing disparities and promoting positive outcomes.

R^2 Values

Furthermore, R^2 values must be considered as a measure of effect size. The primary objective of this dissertation was to evaluate classroom-level predictors (i.e., CLASS Pre-K factors) and how they relate to child outcomes (i.e., change scores from Fall to Spring on Total Protective Factors and Behavioral Concerns). The R^2 values in this study suggest that while the CLASS Pre-K three-factor and bifactor models account for some variance in DECA Total Protective Factors and Behavioral Concerns change scores, a substantial portion of the total outcome variance remains unexplained. Specifically, the three-factor model and bifactor model predictors each accounted for 20% of the variance in change scores from Fall to Spring of Total Protective Factors and 13% of the variance in change scores from Fall to Spring of Behavioral Concerns, which are lower than the expected effect size (R^2 = .30; Wang et al., 2022), but in line with previous literature (Burchinal et al., 2010; Gordon & Peng, 2020; Leyva et al., 2015; Mashburn et al., 2008; Rimm-Kaufman et al., 2009; Zheng et al., 2023). However, the models were more effective at explaining between-classroom variance, capturing 61% of the variance in Total Protective Factors and 56% in Behavioral Concerns. These findings suggest that while classroom-level factors contribute to child outcomes, a sizable portion of the variance exists within classrooms, driven by individual child characteristics. The low overall R^2 values indicate that additional unmeasured factors may influence these outcomes. This further emphasizes the need for both classroom-wide improvements and targeted interventions for individual children, particularly those most at risk. Future research should explore additional predictors, such as teacher experience or family engagement, to better understand the factors shaping these outcomes.

Unmeasured Classroom Level Variables

Several classroom-level factors may influence both CLASS Pre-K and DECA scores that were unmeasured in this dissertation. For instance, children's ages can affect observed behavior and teacher interaction styles, with younger children often requiring more emotional support and behavioral guidance, which may influence Emotional Support scores and DECA socioemotional ratings (Mashburn et al., 2008). Teachers' use of children's home languages may also enhance the quality of interactions by fostering stronger relationships and more effective communication, supporting both higher CLASS Pre-K scores and better socioemotional outcomes (Castro et al., 2011). Additionally, classrooms with a higher number of disruptive or developmentally advanced children may experience more challenges in behavior management and consistency of routines, potentially impacting Classroom Organization scores and child behavior ratings (Blair & Raver, 2015; Pianta et al., 2008b). These contextual variables should be considered when interpreting associations between observed classroom quality and child development outcomes.

Prior Evidence Linking CLASS Pre-K Factor Scores to Child Outcomes

Although the findings in this dissertation were not significant, prior studies have demonstrated statistically significant associations between CLASS Pre-K scores and children's academic and socio-emotional outcomes, supporting the idea that high-quality teacher–child interactions play a meaningful role in early development (Burchinal et al., 2010; Downer et al., 2012; Leyva et al., 2015; Mashburn et al., 2008; Zheng et al., 2023). Emotional Support has been linked to improvements in self-regulation and social competence, while Classroom Organization is often associated with gains in language, literacy, and math (Curby et al., 2009; Rimm-Kaufman et al., 2009; Zheng et al., 2023). Instructional Support has also shown predictive value for language development, although its impact tends to be smaller and more variable across contexts (Burchinal et al., 2010; Gordon & Peng, 2020).

Although statistically significant, the effects reported in prior CLASS Pre-K studies are typically modest in size, ranging from approximately r = .06 to .20 (Burchinal et al., 2010; Gordon & Peng, 2020; Leyva et al., 2015; Mashburn et al., 2008; Rimm-Kaufman et al., 2009; Zheng et al., 2023). The effect sizes observed in this dissertation (r = .20 for Total Protective Factors and r = .13 for Behavioral Concerns) fall well within this range, suggesting alignment with existing research even though the associations were not statistically significant. For instance, Mashburn et al. (2008) found that Emotional Support was linked to improvements in social competence with an effect size of approximately r = .06, while Instructional Support predicted academic outcomes with even smaller effects, around r = .02. Similarly, Burchinal et al. (2010) reported that associations between Instructional Support and children's academic gains rarely exceeded .15. Leyva et al. (2015) also found that the strength of associations between classroom interaction quality and developmental outcomes remained below .20. Downer et al. (2012) observed that teacher–child interactions were more predictive for children at higher developmental risk, although the overall effect sizes remained small (below .20). Collectively, these findings indicate that while CLASS Pre-K can differentiate classroom interaction quality in meaningful ways, its predictive power for child outcomes is limited in magnitude—highlighting the importance of considering additional contextual, instructional, and individual-level variables when evaluating its utility (Gordon & Peng, 2020; Hamre et al., 2014; Zaslow et al., 2010).

This dissertation adds to a growing body of research indicating that the relationship between observed classroom quality and child outcomes is often modest, inconsistent, and context-dependent (Gordon & Peng, 2020; Hamre et al., 2014; Zaslow et al., 2010). Several scholars have noted that while the CLASS framework reliably differentiates classrooms based on interaction quality, its associations with developmental outcomes such as socio-emotional skills and behavior tend to be small and are not consistently replicated across samples (Gordon & Peng, 2020; Hamre et al., 2014; Zaslow et al., 2010). One interpretation is that CLASS Pre-K captures a necessary but insufficient condition for child growth: high-quality interactions may create a foundational environment for learning, but they may not, on their own, produce measurable developmental gains without the presence of other critical inputs such as explicit instructional supports, culturally relevant pedagogy, or individualized intervention strategies (Downer et al., 2012; Hu et al., 2016; Karlsen et al., 2023). Another possibility is that the strength and nature of the associations between CLASS Pre-K factor scores and child outcomes vary across demographic and sociocultural subgroups, such that children in different cultural, linguistic, or developmental contexts respond differently to the same types of observed teacher behaviors (Downer et al., 2012; Hu et al., 2016; Karlsen et al., 2023). In this sense, the null results observed in this dissertation, while potentially influenced by limited statistical power, do not necessarily contradict existing theoretical frameworks. Rather, the null results underscore the need to refine those theories to better account for heterogeneity in both teacher practice and child response (Gordon & Peng, 2020; Hamre et al., 2014; Zaslow et al., 2010).

Threshold Effects and the Predictive Validity of CLASS Pre-K

In addition to demographic and sociocultural considerations, scholars have proposed that CLASS Pre-K factor scores may need to exceed a certain threshold before they begin to yield observable effects on children’s development (Gordon & Peng, 2020). This suggests that low- to mid-range quality classrooms may not provide sufficient support to produce measurable gains in child outcomes, while only classrooms demonstrating consistently high levels of emotionally supportive, well-managed, and instructionally rich interactions are likely to generate meaningful developmental change (Burchinal et al., 2010). If most classrooms in each sample cluster around the mid-range of quality, for example, between 4 and 5 on the CLASS Pre-K factor of Emotional Support—the statistical relationship between interaction quality and child outcomes may be attenuated or altogether undetectable (Zaslow et al., 2010). This limitation becomes particularly relevant in studies conducted within publicly funded programs like Head Start, where structural challenges such as high teacher turnover, variable staff qualifications, and inconsistent implementation of curriculum can constrain the delivery of high-quality interactions (Downer et al., 2012). These contextual barriers may prevent even moderately skilled teachers from reaching the levels of CLASS Pre-K-defined quality needed to impact outcomes, particularly for children with elevated developmental risks or those who require more targeted supports (Hu et al., 2016).

Furthermore, the CLASS Pre-K tool itself may not be sensitive enough to detect qualitative differences in interaction quality within the middle range of scores, which can lead to a ceiling or floor effect in analytic models (Karlsen et al., 2023). A ceiling effect could occur if many classrooms receive high scores—typically in the 6 to 7 range—leaving little room for variability and making it difficult to distinguish truly exceptional interactions from those that are merely above average (Burchinal et al., 2010; Zaslow et al., 2010). Conversely, a floor effect may arise when many classrooms cluster at the low end of the scale, such as scores between 1 and 2, masking important differences in interaction quality among classrooms that all fall below an implicit quality threshold (Gordon & Peng, 2020; Hu et al., 2016). In both cases, the lack of distributional spread reduces the tool's ability to detect meaningful associations between classroom quality and child outcomes, as the scores may not fully capture nuanced variations in teacher–child interactions that matter for development (Hamre et al., 2014; Ng et al., 2021).

These considerations underscore the importance of interpreting CLASS Pre-K factor scores within the broader ecological systems in which teaching and learning occur and suggest that without attending to these threshold and contextual factors, studies may underestimate the true but conditional effects of classroom quality on child development (Hamre et al., 2014).

Summary of CLASS Pre-K's Predictive Utility

In summary, findings from previous studies investigating the use of CLASS Pre-K raise important questions about its predictive validity across diverse educational contexts (Ng et al., 2021). Rather than indicating a failure of the tool itself, previous studies and the null results in this dissertation, suggest that the effectiveness of CLASS Pre-K as a predictor of child outcomes may be contingent on several moderating factors, including sufficient variation in classroom quality, the presence of meaningful thresholds, and alignment between observed teacher–child interactions and children's cultural, linguistic, and developmental characteristics (Gordon & Peng, 2020; Karlsen et al., 2023; Ng et al., 2021). In classrooms where CLASS Pre-K factor scores are clustered in a narrow mid-to-high range, or where the types of interactions valued by the tool diverge from those emphasized in the community, the ability to detect statistically significant associations with child outcomes may be substantially limited (Cadima et al., 2018; Karlsen et al., 2023; Zaslow et al., 2010). These considerations highlight the importance of situating CLASS Pre-K factor scores within a broader ecological framework that accounts for the multiple layers of influence, including teacher, child, classroom, and cultural context, that shape developmental processes (Downer et al., 2012; Hamre et al., 2014). Without attending to such threshold, distributional, and contextual nuances, researchers and practitioners may underestimate the conditional and complex pathways through which classroom quality influences child development (Burchinal et al., 2010; Hu et al., 2016; Zaslow et al., 2010). These findings underscore the need for a more contextually responsive and theoretically nuanced application of CLASS Pre-K in both research and practice (Hamre et al., 2014; Ng et al., 2021; Karlsen et al., 2023).

Limitations

Power and Effect Size.

The primary limitation of this dissertation was its relatively small sample size of 86 classrooms, which restricted statistical power and increased the likelihood of Type II errors—failing to detect real effects that may exist (Hedges & Rhoads, 2010). This is a critical concern given that prior research has consistently shown that the effect sizes linking CLASS Pre-K factor scores to child outcomes are typically small, often ranging from $r = .06$ to .20 (Burchinal et al., 2010; Gordon & Peng, 2020; Leyva et al., 2015; Mashburn et al., 2008; Zheng et al., 2023). For example, Mashburn et al. (2008) analyzed data from over 2,400 children across 671 classrooms, while Downer et al. (2012) utilized data from 2,983 children in 670 classrooms—sample sizes that offer more power than the current study's dataset. Similarly, Leyva et al. (2015) drew from 127 classrooms in Chile, Burchinal et al. (2010) from hundreds of pre-K sites across multiple studies, and Gordon and Peng (2020) relied on two waves of the nationally representative Head Start FACES dataset, encompassing over 800 classrooms. Given these contrasts, the modest sample size in this dissertation may have limited the ability to detect effects of similar magnitude, even when those effects were present.

The effect sizes found in this dissertation for the three-factor and bifactor models of the CLASS Pre-K ($r = .20$ for Total Protective Factors and $r = .13$ for Behavioral Concerns) fall within the range reported by these larger studies, which further suggests that the null findings may reflect power limitations rather than the absence of meaningful relationships (Gordon &

Peng, 2020; Leyva et al., 2015; Mashburn et al., 2008). Future research should consider multi-site collaborations or leveraging multiple national datasets to obtain larger, more diverse samples that are adequately powered to detect small but potentially meaningful effects (Burchinal et al., 2010; Downer et al., 2012; Gordon & Peng, 2020). Doing so would not only enhance statistical reliability but also improve the generalizability of findings across varied program types and demographic contexts (Zheng et al., 2023). In addition, researchers should conduct a priori power analyses based on realistic estimates of expected effect sizes to guide sample size requirements and avoid underpowered studies that may obscure real associations (Hedges & Rhoads, 2010).

Timing of Data Collection.

Additionally, data for this study was collected shortly before the onset of the COVID-19 pandemic over the 2018-2019 school year. Since that time, early childhood education programs have experienced widespread disruptions, including staffing shortages and shifts in classroom dynamics, which may affect the generalizability of these findings (Jalongo, 2021). Research has documented substantial teacher turnover and elevated stress levels among early educators during and after the pandemic, which can influence the consistency and quality of classroom interactions (Jalongo, 2021). Additionally, many children entering early education settings post-pandemic have exhibited delays in communication and social development, likely due to limited peer interaction during extended periods of isolation (Ofsted, 2024). As a result, findings based on pre-pandemic data may not fully reflect the current conditions, challenges, or developmental profiles present in today's early childhood classrooms.

Cultural and Contextual Classroom Features.

Another key limitation is the inability to formally examine model variation by dual language learner (DLL) status due to the small size of this subgroup. Although DLLs comprise a significant portion of the Head Start population, the dataset lacked sufficient power to test for measurement or structural differences by language background (Durán et al., 2016; Espinosa, 2013). Additionally, while prior research suggests that classroom racial composition may influence teacher–child interactions and student outcomes (Downer et al., 2012; Iruka et al., 2010), this variable was not a central focus of the current analysis.

Beyond demographic considerations, substantial variability in classroom-level conditions may also complicate interpretations of CLASS Pre-K scores. Differences in factors such as teacher training, curriculum implementation, and student needs can significantly influence classroom dynamics and child outcomes, yet these contextual features are not directly captured by CLASS Pre-K ratings (LoCasale-Crouch et al., 2007; Downer et al., 2012). For example, two classrooms may receive comparable CLASS Pre-K scores despite differing in their access to instructional materials or their capacity to support children with individualized learning plans. This unmeasured heterogeneity introduces statistical noise and can obscure true associations between observed interaction quality and developmental outcomes (Bassok & Galdo, 2016; Gordon & Peng, 2020). Such concerns are especially salient in resource-constrained Head Start settings, where structural and ecological variation is common and may interact with instructional practices in meaningful ways (Downer et al., 2012; Hu et al., 2016).

Summary of Limitations.

In summary, the lack of statistically significant findings in this dissertation underscores the methodological and conceptual challenges of detecting small but potentially meaningful associations between observed classroom quality and child outcomes in Head Start settings (Hedges & Rhoads, 2010; Gordon & Peng, 2020). While the observed effect sizes fell within the range reported in prior studies using CLASS Pre-K (Burchinal et al., 2010; Gordon & Peng, 2020; Leyva et al., 2015; Mashburn et al., 2008; Zheng et al., 2023), the limited sample size may have reduced the power needed to detect these effects with confidence, raising the possibility that real relationships were masked by statistical noise (Mashburn et al., 2008; Leyva et al., 2015). Moreover, the substantial variability across classrooms, coupled with the narrow scope of the CLASS Pre-K framework, further complicates the interpretation of null results by introducing unmeasured confounding factors that may influence both interaction quality and child development (Downer et al., 2012; Hu et al., 2016). These limitations suggest that the absence of significant findings should not be interpreted as evidence of no relationship, but rather as an indication that current CLASS Pre-K observational measurement tool may not be sufficiently sensitive or inclusive to capture the full complexity of Head Start classrooms, especially with small sample sizes (Gordon & Peng, 2020; Karlsen et al., 2023).

Addressing these limitations is critical for advancing the field's understanding of how teacher–child interactions influence development in under-resourced and culturally diverse early childhood settings (Hu et al., 2016; Ng et al., 2021). Future research should prioritize larger, more statistically robust samples that allow for subgroup analyses and the detection of small effect sizes typical of educational interventions (Hedges & Rhoads, 2010; Burchinal et al., 2010). In addition, there is a need to expand existing measurement frameworks to include dimensions that reflect trauma-informed care, family engagement, and culturally responsive practices, all of which are highly relevant for children served by Head Start but currently underrepresented in the CLASS framework (Downer et al., 2012; Leyva et al., 2015; Karlsen et al., 2023). By refining tools and methods in these ways, researchers can generate more accurate, equitable, and actionable insights into what constitutes quality in early education—and, more importantly, how that quality supports developmental gains for children from historically marginalized communities (Gordon & Peng, 2020; Hamre et al., 2014).

Future Directions: Incorporating Cultural and Contextual Factors

Given the diversity of the Head Start population, future research should prioritize incorporating cultural, linguistic, and socioeconomic factors into evaluations of classroom quality. Many studies informing this field have been conducted in demographically narrow contexts, often with predominantly White or middle-income samples, limiting generalizability to Head Start populations (Downer et al., 2012; Hu et al., 2016; Karlsen et al., 2023; Leyva et al., 2015; Mashburn et al., 2008; Rimm-Kaufman et al., 2009). Even when studies include more diverse samples, subgroup analyses are often underpowered or omitted, making it difficult to assess whether observed patterns hold across racial and linguistic groups (Downer et al., 2012; Leyva et al., 2015). Cultural norms and expectations may also affect how interactions are perceived and rated; for example, behaviors coded as "sensitive" or "warm" in the CLASS Pre-K rubric may reflect dominant cultural values and miss equally supportive but culturally distinct interaction styles (Hu et al., 2016; Karlsen et al., 2023; Ng et al., 2021).

The current dissertation was limited in its ability to examine cultural variation, as the sample size of dual language learners (DLLs)—the most relevant cultural subgroup—was insufficient for testing measurement or structural differences (Durán et al., 2016; Espinosa, 2013). Additionally, although classroom racial composition can shape social dynamics and interactions (Downer et al., 2012; Iruka et al., 2010), it was not a central variable in this analysis. Future studies should ensure adequate representation of DLLs and racially minoritized children to allow for subgroup modeling and the validation of culturally relevant constructs. Such work would also benefit from measuring the ecological conditions that moderate classroom quality, such as teacher training, curriculum adaptation, and culturally responsive practices (Gordon & Peng, 2020; Hamre et al., 2014; LoCasale-Crouch et al., 2007). In sum, addressing these gaps in future research is essential to ensure that classroom quality frameworks like CLASS Pre-K validly and equitably reflect the experiences and developmental needs of the diverse children served by Head Start programs.

In this dissertation, I investigated two key questions regarding CLASS Pre-K: first, whether the traditional three-factor model and the bifactor model offered a valid representation of classroom interactions in Head Start settings; and second, whether these factor scores predicted improvements in children's socio-emotional outcomes over the pre-K year. In line with prior research, the three-factor model demonstrated substantial structural limitations, including poor model fit and high inter-factor correlations, particularly between Emotional Support and Classroom Organization (Gordon & Peng, 2020; Hamre et al., 2014; Karlsen et al., 2023). Additionally, the bifactor model introduced further model estimation issues and conceptual ambiguity, echoing concerns previously raised (Hamre et al., 2014; Ng et al., 2021). Despite these measurement challenges, prior studies have found modest but statistically significant associations between CLASS Pre-K scores and child outcomes, particularly in large, well-powered samples (Burchinal et al., 2010; Downer et al., 2012; Leyva et al., 2015; Mashburn et al., 2008; Zheng et al., 2023). In contrast, the current dissertation found no significant associations between CLASS Pre-K factor scores and gains in Total Protective Factors or reductions in Behavioral Concerns, even though the observed effect sizes ($r = .20$ and $r = .13$) aligned with those previously reported (Gordon & Peng, 2020; Leyva et al., 2015; Mashburn et al., 2008). However, the lack of statistical power makes it difficult to interpret these effect sizes.

While the interpretation of results in this dissertation was hindered by several factors, the findings from previous studies carry particularly important implications for Head Start, a program where CLASS Pre-K scores are directly tied to federal accountability systems and can determine funding, program status, and continuation (Downer et al., 2012; Gordon & Peng, 2020). Because Head Start programs are required to meet certain thresholds on CLASS Pre-K factors to avoid designation as low-performing, the validity and fairness of CLASS Pre-K scores have real consequences for children, families, and educators (Gordon & Peng, 2020; Hamre et al., 2014; Karlsen et al., 2023). However, this dissertation adds to a growing body of evidence suggesting that CLASS Pre-K may not function equally well across all Head Start contexts, especially those serving racially, linguistically, and socioeconomically diverse populations (Hu et al., 2016; Ng et al., 2021). If CLASS Pre-K factor scores do not reliably predict gains in child outcomes, as shown in previous studies, then using these scores as high-stakes indicators of classroom quality may be both misleading and inequitable (Cadima et al., 2018; Gordon & Peng, 2020).

Moreover, when programs are evaluated based on tools that may not be culturally aligned or developmentally comprehensive, they risk being penalized despite engaging in high-quality, community-rooted practices that fall outside the scope of the CLASS Pre-K dimensions (Hu et al., 2016; Karlsen et al., 2023). This could result in funding losses, staff turnover, or reputational harm to programs that are effectively serving their populations, exacerbating inequities rather than alleviating them (Downer et al., 2012; Gordon & Peng, 2020). Therefore, it is critical that policymakers and researchers reevaluate the role of CLASS Pre-K in high-stakes decision-making, especially in settings like Head Start that serve children with the greatest needs (Hamre et al., 2014; Ng et al., 2021). Future iterations of the framework should incorporate culturally relevant indicators and consider alternative models of quality that better reflect the developmental realities and strengths of diverse early learning environments (Cadima et al., 2018; Leyva et al., 2015).

Missing Data

Further complicating the interpretation of these null findings is the issue of missing child outcome data, which poses a significant threat to the validity of statistical inferences in educational research (Enders, 2010; Sabol et al., 2020). In this study, missingness was not randomly distributed across classrooms or predictor quartiles, suggesting that some classrooms may systematically differ in ways that influence both assessment participation and observed outcomes (Schafer & Graham, 2002; Sabol et al., 2020). For example, classrooms serving children with higher behavioral needs, limited teacher availability, or greater structural disadvantage may be more prone to missing data due to logistical barriers to data collection (Downer et al., 2012; Hu et al., 2016). If these characteristics also relate to classroom quality or development, then the missing data may introduce selection bias, potentially distorting associations between CLASS Pre-K scores and child outcomes (Enders, 2010; Schafer & Graham, 2002).

To address these concerns related to missing data, researchers increasingly recommend the use of multiple imputation techniques, which estimate missing values using information from observed data to reduce bias and improve power (Enders, 2010; Schafer & Graham, 2002, van Buuren, 2018). Multiple imputation is particularly effective when missingness is "missing at random" (MAR), meaning that the probability of a value being missing is related to observed data but not to the missing data itself (Harel & Zhou, 2007). In other words, once relevant observed variables are accounted for, the missingness does not depend on the unobserved values (Harel & Zhou, 2007). This assumption is less restrictive than missing completely at random (MCAR) and is often more realistic in educational research settings (Sabol et al., 2020). When MAR holds, multiple imputation has been shown to produce more accurate and robust parameter estimates in clustered educational datasets (Harel & Zhou, 2007; Sabol et al., 2020). Because missing Spring scores were plausibly related to observed Fall scores, which were included in the imputation model, the assumption of missing at random (MAR) is reasonable in this context (Harel & Zhou, 2007; Sabol et al., 2020). For these reasons, Fall scores were used as predictors in the imputation of missing Spring outcome scores to help account for systematic differences between observed and missing cases.

Additionally, sensitivity analyses—such as comparing results across complete-case analyses and imputed datasets—can provide insight into the stability of findings under different assumptions about the missing data mechanism (Jakobsen et al., 2017; Schafer & Graham, 2002). These approaches are especially important in Head Start and other high-need early education settings, where missing data are often entangled with structural inequities and should not be assumed to occur randomly (Downer et al., 2012; Sabol et al., 2020). Without adequate handling of missingness, findings may underestimate or overestimate the true associations between classroom quality and child development, limiting both scientific validity and policy relevance (Enders, 2010; Hu et al., 2016). For these reasons, it is recommended that missing data is thoroughly documented and handled appropriately in future studies.

Theoretical Limitations of CLASS Pre-K

Aside from implications of missing data, limitations regarding applying the CLASS Pre-K theoretical framework to Head Start populations exist. CLASS Pre-K was developed to provide a standardized, research-based measure of teacher–child interactions grounded in the Teaching Through Interactions framework (Pianta et al., 2008b; Hamre et al., 2014). Theoretically, CLASS Pre-K is intended to assess emotionally supportive, well-organized, and instructionally rich environments that foster young children's cognitive and socio-emotional development (Pianta & Hamre, 2009). However, findings from previous studies challenge key aspects of CLASS Pre-K's legitimacy. The poor fit of both the three-factor and bifactor models raises questions about the structural soundness of the tool in diverse classroom settings (Hamre et al., 2014; Gordon & Peng, 2020; Ng et al., 2021). Furthermore, the absence of statistically significant associations with low effect sizes between CLASS Pre-K scores and gains in child outcomes suggests a disconnect between observed teacher–child interactions and actual developmental progress (Gordon & Peng, 2020; Hu et al., 2016).

These findings are not isolated. Previous studies have also reported only modest and inconsistent associations between CLASS Pre-K scores and child outcomes, particularly in programs serving racially, linguistically, and socioeconomically diverse populations (Burchinal et al., 2010; Downer et al., 2012; Gordon & Peng, 2020; Leyva et al., 2015; Mashburn et al., 2008). While CLASS Pre-K reliably distinguishes between higher- and lower-quality interactions, its predictive validity is limited and may be highly context-dependent (Gordon & Peng, 2020; Zaslow et al., 2010). Moreover, concerns about cultural misalignment and underrepresentation suggest that CLASS Pre-K may not capture the full range of high-quality practices needed in settings like Head Start, where trauma-informed care, bilingual scaffolding, and family engagement are often critical (Hu et al., 2016; Karlsen et al., 2023). Taken together, the theoretical intentions of CLASS Pre-K—to serve as a fair and actionable indicator of teaching quality—are undermined by mounting empirical evidence that its structural design, cultural relevance, and predictive utility fall short in the very contexts where it is most consequential. If CLASS Pre-K is to remain a centerpiece of early childhood quality measurement, it must be meaningfully revised to better reflect the complexity and diversity of contemporary classrooms (Cadima et al., 2018; Ng et al., 2021).

Structural Limitations of CLASS Pre-K in Diverse Head Start Classrooms

Several studies have identified critical limitations in the structural integrity of the CLASS Pre-K measurement model, particularly when used in culturally and linguistically diverse Head Start settings (Gordon & Peng, 2020; Ng et al., 2021). Poor model fit has been reported in both the traditional three-factor and bifactor models, suggesting that the latent structure does not adequately represent observed patterns of teacher–child interaction (Brown, 2015; Hu & Bentler, 1999). One commonly reported issue is the extremely high correlation between Emotional Support and Classroom Organization in the three-factor model (e.g., $r \approx .96$), which raises concerns about discriminant validity and the distinctiveness of these domains (Gordon & Peng, 2020; Li et al., 2020). High intercorrelations among latent factors can undermine the validity of interpreting domain scores as empirically separate constructs (Bihler et al., 2018; Cadima et al., 2018), which is especially problematic in Head Start contexts where differentiated feedback is critical (Downer et al., 2012; Hu et al., 2016). Bifactor models also exhibit structural weaknesses, including negative residual variances and weak item loadings ($\lambda < .40$), indicating problems with model specification and reliability (Brown, 2015; Karlsen et al., 2023). Certain dimensions—such as Negative Climate and Instructional Support—often perform inconsistently across studies, failing to load reliably or show adequate variability, particularly in low-income or multilingual classrooms (Hamre et al., 2014; Pakarinen et al., 2010; Cadima et al., 2018; Leyva et al., 2015). These limitations call into question the generalizability and fairness of the CLASS Pre-K in diverse educational settings and underscore the need for culturally responsive revisions to ensure psychometric and contextual validity in high-stakes applications (Gordon & Peng, 2020; Karlsen et al., 2023; Ng et al., 2021).

Head Start Designation Renewal System

The continued use of CLASS Pre-K in the Designation Renewal System (DRS) raises significant concerns given its structural limitations and modest predictive validity for child outcomes (Gordon & Peng, 2020; Hamre et al., 2014). Under current DRS policy, Head Start programs with CLASS Pre-K scores below predetermined cutoffs are flagged for possible recompetition, making the tool a high-stakes component of program accountability (Gordon & Peng, 2020). However, the review of prior research included in this dissertation highlights substantial weaknesses in the CLASS Pre-K measurement model that undermine its legitimacy for high-stakes applications (Downer et al., 2012; Gordon & Peng, 2020; Ng et al., 2021).

Empirical studies have consistently identified poor model fit challenging the assumption that these domains capture distinct constructs (Gordon & Peng, 2020; Hamre et al., 2014; Li et al., 2020; Ng et al., 2021). These structural issues were replicated in the current study, further casting doubt on the internal consistency and construct validity of the CLASS framework in diverse classroom settings (Karlsen et al., 2023; Ng et al., 2021). When factor boundaries are not psychometrically supported, composite CLASS Pre-K dimension scores may fail to offer accurate or meaningful information about overall classroom quality (Cadima et al., 2018; Hamre et al., 2014).

Moreover, the CLASS Pre-K tool has demonstrated only modest associations with child outcomes, with effect sizes typically ranging from r = 06 to .20 across studies, raising questions about the tool's utility for identifying effective classrooms in high-need settings like Head Start (Gordon & Peng, 2020; Zheng et al., 2023). Prior studies have also noted that associations between CLASS Pre-K scores and child outcomes are often stronger for children at higher risk, but remain small overall (Downer et al., 2012; Zaslow et al., 2010).

These findings challenge the core assumption embedded in the DRS: that CLASS Pre-K scores are valid proxies for program effectiveness or children's developmental progress (Hamre et al., 2014; Gordon & Peng, 2020). Without strong, consistent links between CLASS Pre-K scores and child outcomes, using the tool as a gatekeeper for federal funding may produce inequitable consequences, particularly for programs serving children from racially, linguistically, and socioeconomically marginalized communities (Downer et al., 2012; Hu et al., 2016). This concern is especially salient given research showing that CLASS Pre-K dimensions—such as *Teacher Sensitivity* or *Behavior Management*—may be enacted and interpreted differently across cultural contexts, potentially leading to biased or incomplete assessments (Bihler et al., 2018; Leyva et al., 2015; Ng et al., 2021).

Several studies have also raised concerns about the CLASS framework's limited sensitivity to culturally responsive teaching practices, which are critical in settings like Head Start where educators are tasked with supporting children who may be navigating trauma, linguistic barriers, or systemic inequities (Downer et al., 2012; Gordon & Peng, 2020; Hu et al., 2016; Karlsen et al., 2023). If CLASS Pre-K fails to capture these contextually embedded practices, programs may receive low scores not because they are ineffective, but because the measurement tool does not reflect the full range of high-quality teaching relevant to their communities (Gordon & Peng, 2020; Ng et al., 2021).

Given these limitations, relying on CLASS Pre-K scores to determine whether Head Start programs retain funding under the DRS is highly problematic (Gordon & Peng, 2020; Hamre et al., 2014; Karlsen et al., 2023). This approach risks penalizing programs that excel in ways not recognized by CLASS Pre-K metrics, while potentially rewarding others whose scores reflect alignment with dominant cultural norms rather than true effectiveness (Leyva et al., 2015; Ng et al., 2021). Furthermore, such misuse of data can undermine trust in early childhood accountability systems and reduce the credibility of CLASS Pre-K as a tool for continuous quality improvement (Cadima et al., 2018; Gordon & Peng, 2020).

To ensure that the DRS promotes equitable and evidence-based decision-making, future policy should reconsider the centrality of CLASS Pre-K scores and explore multi-dimensional approaches that incorporate broader indicators of quality, including culturally responsive instruction, trauma-informed care, and family engagement (Downer et al., 2012; Hu et al., 2016; Karlsen et al., 2023). Until the CLASS Pre-K framework is refined to address these conceptual and measurement shortcomings, its use in the DRS should be critically re-evaluated to avoid unintended harm to the very communities Head Start is intended to support (Gordon & Peng, 2020; Ng et al., 2021).

Recommendations for Improving Measurement of Classroom Quality in Head Start

To capture the complexity of classroom quality more accurately in Head Start settings, revisions to the CLASS framework are urgently needed to address its current conceptual, structural, and cultural limitations (Gordon & Peng, 2020; Ng et al., 2021). Existing dimensions such as Emotional Support, Classroom Organization, and Instructional Support reflect foundational components of teacher–child interaction, but they may not fully capture the types of relational and instructional practices most relevant for children in high-poverty, culturally diverse settings like Head Start (Downer et al., 2012; Hu et al., 2016; Leyva et al., 2015). For example, current CLASS Pre-K dimensions often overlook culturally responsive practices—such as storytelling, home language use, or relational approaches to discipline—that are critical for engaging children from racially and linguistically minoritized backgrounds (Cadima et al., 2018; Ng et al., 2021). Incorporating such practices into revised CLASS Pre-K dimensions would enhance the tool's cultural validity and ensure a more accurate representation of teaching quality across different communities (Gordon & Peng, 2020; Leyva et al., 2015).

Moreover, researchers have argued for expanding the CLASS framework to include constructs such as trauma-informed care and family engagement, which are particularly important in Head Start classrooms that serve children exposed to chronic stress, instability, and systemic disadvantage (Hu et al., 2016; Karlsen et al., 2023). These practices go beyond traditional Emotional Support to include regulatory, reparative, and resilience-promoting strategies, which are especially impactful in high-need contexts but currently unmeasured by CLASS Pre-K (Downer et al., 2012; Gordon & Peng, 2020). Integrating observational items that assess teacher responsiveness to children's emotional cues, culturally affirming communication styles, and inclusive family engagement would improve the construct validity and practical utility of the tool (Karlsen et al., 2023; Ng et al., 2021).

Additionally, measurement models should account for variability in classroom context by incorporating flexible scoring systems or adaptive observational rubrics that reflect both developmental appropriateness and cultural alignment (Cadima et al., 2018; Gordon & Peng, 2020). Given the structural challenges and diversity within Head Start, such as multilingualism, varying teacher training levels, and heterogeneous child needs, a one-size-fits-all measurement model may fail to differentiate between truly high- and low-quality interactions (Downer et al., 2012; Hu et al., 2016). Including supplementary tools—such as companion rubrics for family engagement or trauma-informed practice—could further enhance the ecological validity of assessments and inform more equitable coaching and support strategies (Gordon & Peng, 2020; Leyva et al., 2015).

In sum, to strengthen both the scientific rigor and applied value of classroom observation in Head Start, the CLASS Pre-K framework should be revised to incorporate culturally grounded, trauma-responsive, and family-centered indicators of quality, and should adopt more flexible measurement models that reflect the realities of diverse early childhood contexts (Downer et al., 2012; Hu et al., 2016; Karlsen et al., 2023; Ng et al., 2021).

Policy Recommendations

Given the documented measurement limitations and context-specific challenges associated with the CLASS Pre-K tool, policymakers should adopt a more nuanced and equitable approach to evaluating classroom quality in Head Start programs (Gordon & Peng, 2020; Hamre et al., 2014; Ng et al., 2021). The current reliance on CLASS Pre-K scores within the Designation Renewal System (DRS) for making high-stakes decisions—such as program funding and continuation—places undue weight on a tool that may not fully or fairly capture the complexity of teaching quality in culturally and linguistically diverse settings (Downer et al., 2012; Hu et al., 2016). Evidence from previous well-powered studies shows that CLASS Pre-K scores do not consistently predict gains in child outcomes, particularly in high-need populations like those served by Head Start (Leyva et al., 2015; Mashburn et al., 2008; Zheng et al., 2023). This calls into question the appropriateness of using CLASS Pre-K as a determinant of program effectiveness (Gordon & Peng, 2020; Zaslow et al., 2010).

To promote greater accuracy and fairness, policymakers should require that CLASS Pre-K scores be supplemented with additional sources of data—such as measures of family engagement, child progress monitoring, teacher professional development, and contextual indicators of structural quality (Downer et al., 2012; Karlsen et al., 2023). These complementary metrics would provide a more comprehensive picture of program performance and help ensure that evaluations reflect the lived experiences of children and families in under-resourced communities (Cadima et al., 2018; Ng et al., 2021). Relying on observational scores that may be misaligned with community values or culturally variable practices risks penalizing programs that are, in fact, delivering high-quality instruction in ways not captured by the CLASS framework (Gordon & Peng, 2020; Leyva et al., 2015).

If left unaddressed, these measurement limitations can lead to unintended consequences, such as the misallocation of resources, unjust comparisons among centers, and diminished trust in federal accountability systems (Gordon & Peng, 2020; Hamre et al., 2014; Hu et al., 2016). Programs serving the highest-need populations may be disproportionately penalized despite making strong efforts to deliver responsive and culturally attuned instruction (Downer et al., 2012; Gordon & Peng, 2020). To mitigate these risks, federal and state agencies should revise DRS criteria to include multiple indicators of classroom quality, prioritize equity in evaluation, and invest in tools that reflect the strengths and needs of diverse early childhood settings (Karlsen et al., 2023; Ng et al., 2021). By implementing these changes, policymakers can better ensure that classroom quality assessments support—not hinder—the overarching goals of Head Start: to provide inclusive, high-quality learning environments that promote positive outcomes for all children (Downer et al., 2012; Hu et al., 2016).

Summary of Recommendations

To ensure more equitable and accurate evaluation in the Designation Renewal System (DRS), future research should prioritize validating classroom quality measures in the specific cultural and structural contexts where they are most often applied, such as Head Start classrooms (Downer et al., 2012; Hu et al., 2016). Given that the CLASS Pre-K framework demonstrates limited structural validity and predictive power in diverse, high-need settings, it is imperative that researchers conduct further confirmatory and exploratory factor analyses in large, demographically representative samples, including classrooms serving racially minoritized and dual language learners (Gordon & Peng, 2020; Karlsen et al., 2023; Ng et al., 2021). These studies should assess the consistency of CLASS Pre-K factor structures across subgroups to determine whether the tool meets criteria for measurement invariance—a key requirement for fair use in accountability systems like DRS (Bihler et al., 2018; Hu et al., 2016).

Additionally, future work should explore the development and validation of expanded observational frameworks that incorporate context-relevant practices, such as trauma-informed care, culturally responsive pedagogy, and family engagement (Downer et al., 2012; Karlsen et al., 2023; Leyva et al., 2015). These dimensions are essential for understanding quality in classrooms that serve children exposed to poverty, racism, displacement, and other systemic barriers, yet they are not currently emphasized within the CLASS Pre-K rubric (Hu et al., 2016; Ng et al., 2021). New observational tools or companion metrics should be piloted alongside CLASS Pre-K to test whether they provide additional explanatory power for child outcomes—particularly for socio-emotional growth and behavioral regulation, which are core targets of Head Start (Cadima et al., 2018; Gordon & Peng, 2020).

Longitudinal studies that track how classroom interactions evolve over time and how these changes relate to child growth would also strengthen the evidence base guiding DRS policy (Hamre et al., 2014; Pianta & Hamre, 2009). These studies should investigate potential threshold effects, exploring whether only exceptionally high-quality interactions yield observable gains, and whether moderate or low scores on CLASS Pre-K fail to differentiate instructional effectiveness in high-need settings (Burchinal et al., 2010; Zaslow et al., 2010). Future analyses should also examine the predictive utility of CLASS Pre-K within multi-method designs, incorporating both quantitative and qualitative data, to capture teacher practices that may be overlooked in standard observation cycles (Gordon & Peng, 2020; Karlsen et al., 2023).

Finally, researchers should investigate how DRS decisions affect programs on the ground, including whether centers that fail to meet CLASS Pre-K factor score thresholds are indeed offering lower-quality experiences, or whether they are disproportionately penalized due to demographic factors unrelated to instructional quality (Downer et al., 2012; Gordon & Peng, 2020). Equity audits and policy simulations could be used to model alternative accountability systems that weigh multiple indicators, such as child growth metrics, family engagement data, and professional development investments, to paint a more comprehensive and just picture of program quality (Leyva et al., 2015; Ng et al., 2021). By pursuing these lines of inquiry, the field can move toward a more valid, culturally attuned, and policy-relevant approach to evaluating early childhood education quality in programs like Head Start (Hamre et al., 2014; Karlsen et al., 2023).

FINAL CONCLUSIONS

This dissertation critically examined the measurement structure and predictive validity of the CLASS Pre-K framework in Head Start settings, highlighting important limitations and implications for research, practice, and policy (Downer et al., 2012; Hamre et al., 2014; Ng et al., 2021). I found that across studies the three-factor model and the bifactor model exhibited poor model fit, high inter-factor correlations, and inconsistent item loadings, raising concerns about their structural validity in diverse early childhood classrooms (Gordon & Peng, 2020; Karlsen et al., 2023; Ng et al., 2021). In addition to factor structure limitations, a review of previous research revealed inconsistent associations and modest effect sizes between CLASS Pre-K and child outcomes (Burchinal et al., 2010; Downer et al., 2012; Gordon & Peng, 2020; Hu et al., 2016; Leyva et al., 2015; Mashburn et al., 2008; Zheng et al., 2023).

Findings from previous studies reflect deeper conceptual limitations within the CLASS framework—particularly its narrow scope and cultural assumptions (Hu et al., 2016; Karlsen et al., 2023). In high-need environments like Head Start, classroom quality encompasses more than emotionally supportive and instructionally rich interactions; it includes trauma-informed care, culturally responsive teaching, and strong teacher–family partnerships, all of which are underrepresented in CLASS Pre-K metrics (Downer et al., 2012; Leyva et al., 2015; Ng et al., 2021). This mismatch between what is measured and what matters may help explain why CLASS Pre-K scores did not predict gains in socio-emotional development in this dissertation (Gordon & Peng, 2020; Karlsen et al., 2023).

Given these findings, it is critical to re-evaluate the role of CLASS Pre-K in high-stakes accountability systems such as the Designation Renewal System (DRS), which relies heavily on

CLASS Pre-K scores to make funding and program continuation decisions (Bassok & Galdo, 2016; Downer et al., 2012). When measurement tools lack cultural sensitivity, conceptual clarity, and predictive power, their use in high-stakes decisions can result in misaligned coaching, biased evaluations, and inequitable resource distribution (Gordon & Peng, 2020; Hu et al., 2016). In this context, continued reliance on unmodified CLASS Pre-K scores risks penalizing programs that serve children most in need while failing to capture the full range of high-quality teaching practices present in diverse classrooms (Karlsen et al., 2023; Ng et al., 2021).

To strengthen the validity, fairness, and utility of classroom quality assessment, this dissertation recommends revising the CLASS Pre-K framework to better reflect the interactional, cultural, and structural realities of Head Start environments (Cadima et al., 2018; Hu et al., 2016). Future iterations should incorporate indicators of trauma-informed care, culturally relevant pedagogy, and family engagement—elements that are essential for supporting child development in historically marginalized communities but absent from the current framework (Downer et al., 2012; Leyva et al., 2015; Ng et al., 2021). Mixed-methods studies, subgroup-specific validations, and participatory research with educators and families will be essential to building tools that are both psychometrically sound and contextually responsive (Gordon & Peng, 2020; Karlsen et al., 2023).

In conclusion, while CLASS Pre-K offers a valuable foundation for observing and understanding teacher–child interactions, its current structure and application fall short of meeting the complex needs of diverse early learning environments like Head Start (Gordon & Peng, 2020; Hamre et al., 2014; Ng et al., 2021). The findings from studies reviewed in this dissertation underscore the urgent need for theoretical refinement, cultural adaptation, and policy reform to ensure that classroom quality measurement genuinely supports equity, learning, and developmental gains for all children in Head Start.

Table 1

Descriptions of the Ten Dimensions and Three Domains of the three-factor model of CLASS Pre-K

Dimension	Description	Examples
Emotional Support Domain		
Positive Climate	Considers the comfort, warmth, and respect displayed in teachers' and children' interactions with one another and the degree to which they display enjoyment during learning activities.	Teachers and children have positive emotions, communication, and respectful interactions.
Negative Climate	Reflects the level of expressed negativity such as anger, hostility, or aggression demonstrated by teachers and/or children.	Teachers and children display sarcasm, disrespect, and negativity. Teachers only: Punitive control
Teacher Sensitivity	Encompasses teachers' awareness of and responsivity to children' individual academic and social-emotional needs.	Teachers are aware of and respond. to children who need extra support and address problems promptly.
Regard for Child Perspectives	The degree to which teachers' interactions with children emphasize children' interests and ideas and promote child autonomy rather than being very teacher-directed.	Teachers promote and incorporate children's ideas, support independence, and create a child-centered classroom.
Instructional Support Domain		
Concept Development	The degree to which instructional discussions and activities promote children' higher-order thinking skills versus rote learning.	Teachers are clear, integrate knowledge, make learning meaningful, and encourage deep thinking.
Quality of Feedback	Involves how teachers provide feedback focused on expanding children's learning and understanding versus correctness.	Teachers scaffold learning, build on children's responses, and provide encouragement and affirmation.

Language Modeling	Involves teachers using language-facilitation techniques, including self and parallel talk, open-ended questions, repetition and extension, and use of advanced vocabulary.	Teachers promote and encourage talking and use advanced language.
Classroom Organization Domain		
Behavior Management	Reflects teachers' use of effective methods to prevent and redirect misbehavior by communicating clear behavioral expectations and minimizing time spent reacting to issues.	Teachers are proactive, establish clear expectations, and effectively redirect misbehavior.
Productivity	Considers how well teachers manage instructional time, transitions, and routines so that children have maximal opportunities to learn.	Teachers maximize learning time, establish clear routines, and make transitions brief and engaging.
Instructional Learning Formats	The degree to which teachers maximize children' engagement by providing clear learning objectives, interesting materials, and facilitation.	

Note: Classroom Assessment Scoring System and CLASS are registered trademarks of Teachstone, Inc. (Teachstone.com)

Table 2

The Bifactor Model of CLASS Pre-K (Hamre et al., 2013)

General Domain	Responsive Teaching	
General Only Dimensions*	Teacher Sensitivity	
	Regard for Child Perspectives	
	Instructional Learning Formats	
Specific Domains	Proactive Management and Routines	Cognitive Facilitation
Specific Dimensions	Positive Climate	Concept Development
	Negative Climate	Quality of Feedback
	Behavior Management	Language Modeling
	Productivity	

**Note: Responsive Teaching is reflected in all 10 dimensions.*

Table 3

Fit Statistics of Previous CFAs of CLASS Pre-K

Paper	Fit Results: One factor	Fit Results: Two factors	Fit Results: Three-factors	Fit Results: Bifactor
Bihler et al., 2018	$X^2(df)$= 147.47(35) p <.05 CFI= .65 RMSEA= .14 TLI= .55	$X^2(df)$= 81.09 (33) p <.05 CFI= .85 RMSEA= .09 TLI= .80	**$X^2(df)$= 48.15(29)** **p <.05** **CFI= .94** **RMSEA= .06** **TLI= .91**	$X^2(df)$= 67.64(26) p <.05 CFI= .87 RMSEA= .10 TLI= .78
Cadima et al., 2018	$X^2(df)$= 532.59(27) p <.001 SRMR= .16 CFI= .64 RMSEA= .12 TLI= .51	$X^2(df)$= 120.50(26) p <.001 SRMR= .07 CFI= .93 RMSEA= .14 TLI= .91	**$X^2(df)$= 74.67(24)** **p <.001** **SRMR= .05** **CFI= .96** **RMSEA= .11** **TLI= .95**	
Downer et al, 2012			**$X^2(df)$= 174.6(59)** **p <.01** **CFI= .97** **RMSEA= .09** **TLI= .95**	
Gordon and Peng, 2020 (FACES 2009)	CFI= .71 NNFI= .62 RMSEA= .18	CFI= .91 NNFI= .88 RMSEA= .10	**CFI= .94** **NNFI= .92** **RMSEA= .08**	CFI= .96 NNFI= .92 RMSEA= .08
Gordon and Peng, 2020 (FACES 2014)	CFI= .64 NNFI= .53 RMSEA= .20	CFI= .94 NNFI= .91 RMSEA= .90	CFI= .94 NNFI= .91 RMSEA= .09	**CFI= .96** **NNFI= .94** **RMSEA= .08**
Hamre et al., 2013	$X^2(df)$= 993(65) p <.001 CFI= .78 TLI= .74 RMSEA= .05	$X^2(df)$= 849(64) p <.001 CFI= .82 TLI=.78 RMSEA= .05	**$X^2(df)$= 728(62)** **p <.001** **CFI= .95** **TLI=.80** **RMSEA= .05**	

	SRMR= .05	SRMR=.05	**SRMR=.05**	
Hamre et al., 2014	CFI= .94 RMSEA= .13 SRMR= .07	CFI= .92 RMSEA= .14 SRMR= .07	CFI= .93 RMSEA= .13 SRMR= .08	**CFI= .96** **RMSEA= .11** **SRMR= .04**
Hu et al., 2016	$X^2(df)$= 546.55(50) p <.001 CFI= .61 RMSEA= .24 TLI= .65	$X^2(df)$= 227.00(48) p <.001 CFI= .86 RMSEA= .14 TLI= .87	**$X^2(df)$= 161.89(49)** **p <.001** **CFI= .91** **RMSEA= .11** **TLI= .92**	**$X^2(df)$ = 147.21(44)** **p <.001** **CFI= .92** **RMSEA= .11** **TLI= .92**
Karlsen et al., 2023			**CFI= .93** **TLI= .89** **RMSEA= .11** **SRMR= .07**	
Leyva et al., 2015			**$X^2(df)$= 58.28(28)** **p <.001** **CFI= .93** **TLI= .88** **RMSEA= .11** **SRMR= .06**	
Ng et al., 2021	AIC: 3307.92 BIC: 3404.36	AIC: 3101.63 BIC: 3201.29	AIC: 3083.23 BIC: 3189.33	**AIC: 2972.84** **BIC: 3098.22**
Pakarinen et al., 2010	$X^2(df)$= 116.27(27) p = .000 CFI= .83 TLI= .78 RMSEA= .26 SRMR= .06		**X^2 (df)= 45.16(23)** **p = .004** **CFI= .96** **TLI= .94** **RMSEA= .14** **SRMR= .04**	

Table 4

Variables in Previous MLMs with Scores from the Traditional Three-factor Structure of CLASS Pre-K as Predictors

Study	Predictors	Covariates	Outcomes	Missing Data	Results
Burchinal et al., 2010	Emotional Support Instructional Support (average dimension scores)	Fall pre-test scores Gender Race/Ethnicity Time between assessments Home language Site identifier Type of program (Head Start or other) Program setting (school or not)	Language skills Literacy skills Math skills Social competence Behavior problems	Multiple imputation	Instructional Support predicted language and literacy gains in classrooms with moderate to high-quality scores. Emotional Support predicted social competence gains and reduced behavior problems in classrooms with high quality scores. Threshold effect- more evident impacts of quality on outcomes in higher quality classrooms. No significant thresholds.
Curby et al., 2009a	Emotinal Support Instructional Support Classroom Organization	Fall pre-test scores Child age, gender, ethnicity, family poverty status, maternal education Time between assessments	Receptive vocabulary Math skills Social competence	Only used complete data for outcomes	Emotional Support was associated with higher social competence in classrooms with high scores. There were significant differences in gains between highest quality and lowest quality classrooms with small effect sizes. The dimension of Concept Development showed gains in receptive vocabulary and math skills.

Curby et al., 2009b	Emotional Support Instructional Support Classroom Organization	Fall pre-test scores Child gender, SES, maternal education	Letter-word identification Phonological awareness Math skills	Only used complete data	Emotional Support was positively associated with phonological awareness. Instructional Support benefited children with lower initial letter-word identification scores. Classroom Organization benefited children with lower math scores.
Downer et al., 2012	Emotional Support Instructional Support Classroom Organization	Child gender, age, race, maternal education, SES, language spoken at home, test interval, Fall scores. Teacher degree, teacher-child ratio, full day status, % of dual language learners in classroom.	Letter naming Math skills Language and literacy skills Social competence Problem behaviors	Full information maximum likelihood	Emotional Support predicted social competence and letter naming. Classroom Organization predicted math, social competence, problem behaviors, and letter naming. Instructional Support predicted language/literacy, math, and letter naming.
Gordon & Peng, 2020	Emotional Support Instructional Support Classroom Organization Emotional Support and Classroom Organization combined	Child (all pre-test scores, gender, race, poverty status, disability status, maternal education, age, and time between assessments(Classroom level (teacher education, years of experience,	Letter-word identification Math skills Receptive vocabulary Social skills Behavior problems Inhibitory control Behavior during assessments	Multiple imputation	Small and typically non-significant associations with gains children's academic and social skills. 12 out of 240 regressions were significant, did not replicate across samples, and effects were very small.

		number of children, classroom composition)			
Hamre et al., 2014	Emotional Support Instructional Support Classroom Organization Responsive Teaching Proactive Management and Routines Cognitive Facilitation	Child (age, gender, maternal education, time between assessments, Fall scores), and classroom (teacher education, teacher years of experience, classroom setting, literacy curriculum)	Vocabulary Phonological awareness Working memory Inhibitory control Closeness to teacher and conflict in teacher-child relationships	Full information maximum likelihood	Responsive Teaching predicted gains in literacy and language, working memory, and reductions in teacher-reported conflict. Cognitive Facilitation predicted gains in literacy and language. Proactive Management and Routines predicted gains in inhibitory control. Emotional Support negatively associated with gains in inhibitory control. Instructional Support predicted gains in literacy and language and was associated with closer teacher-child relationships. Classroom Organization predicted gains in inhibitory control, working memory, and literacy. Small effects for all associations.
Hatfield et al., 2016	Emotional Support Instructional Support Classroom Organization	Fall pre-test scores Gender Race/Ethnicity Maternal education Home language Site location	Phonological awareness Inhibitory control		Emotional Support above the threshold of 6 was a strong predictor of phonological awareness and inhibitory control.

					Associations are close to 0 below threshold.
Leyva et al., 2015	Emotional Support Instructional Support Classroom Organization	Child age and gender. Teacher age, experience, postgrad education, years of experience.	Vocabulary Letter-Word identification Writing skills Early numeracy Inhibitory control Cognitive flexibility	Multiple imputation	Instructional Support predicted gains in writing skills and inhibitory control, and stronger positive associations with higher levels of quality for language, writing skills, and math skills. Classroom Organization predicted gains in math skills, and stronger negative associations with lower levels of inhibitory control at the lower levels of quality.
Mashburn et al., 2008	Emotional Support Instructional Support	Child pre-test scores, gender, race/ethnicity, maternal education, poverty status. State where program is located.	Receptive vocabulary Expressive language Phonological awareness Early numeracy Letter naming Social competence Problem behaviors	Multiple imputation	Instructional Support was positively associated with gains in all five academic outcomes tested. Emotional Support was positively associated with social competence, and negatively associated with problem behaviors.
Rimm-Kaufman et al., 2009	Emotional Support Instructional Support Classroom Organization	Child gender, pre-K enrollment status, single parent-status,	Behavioral self-control Cognitive self-control	Listwise deletion	Emotional Support had no significant associations. Instructional Support was negatively associated with

	poverty status, maternal education.	Positive work habits Time off-task Engagement in learning		cognitive self-control and positive work habits. Classroom Organization predicted positive gains in all outcomes.

Table 5

Descriptive Statistics for All Variables

	n	Mean	SD	Min	Max
Variables in Confirmatory Factor Analyses					
Positive Climate	86	6.44	0.61	4.25	7.00
Negative Climate	86	1.07	0.18	1.00	1.75
Teacher Sensitivity	86	6.09	0.70	4.25	7.00
Regard for Child Perspectives	86	5.79	0.79	3.00	7.00
Concept Development	86	3.14	0.83	1.00	5.00
Quality of Feedback	86	3.36	1.04	1.50	6.25
Language Modeling	86	3.91	0.98	1.75	6.50
Behavior Management	86	5.83	0.90	3.25	7.00
Productivity	86	5.96	0.84	3.50	7.00
Instructional Learning Formats	86	5.47	0.91	3.25	7.00
Variables in Multilevel Models					
Emotional Support	1418	0.02	0.95	-2.96	1.47
Instructional Support	1418	0.02	0.95	-2.54	1.55
Classroom Organization	1418	0.02	0.93	-1.89	2.08
Responsive Teaching	1418	0.02	0.96	-2.66	1.50
Proactive Management and Routines	1418	0.00	0.75	-2.20	1.80
Cognitive Facilitation	1418	0.01	0.92	-2.03	1.95
Total Protective Factors Fall Scores	1418	48.98	10.20	20.93	72.00
Total Protective Factors Spring Scores	1418	50.37	9.55	17.84	93.00
Total Protective Factors Change Scores	1418	1.38	11.33	-35.60	35.74
Behavioral Concerns Fall Scores	1418	49.81	9.81	24.70	75.42
Behavioral Concerns Spring Scores	1418	50.88	9.60	18.84	83.94
Behavioral Concerns Change Scores	1418	1.07	11.06	-41.16	39.73
Total Protective Factors Fall Classroom Mean	1418	-0.09	6.30	-11.68	16.67
Behavioral Concerns Fall Classroom Mean	1418	-0.02	4.61	-13.76	9.67

Note: All CFA variables consist of observer-averaged item scores, representing the most granular unit of measurement available for analysis, and were collected in the spring. Standardized variables include Emotional Support, Instructional Support, Classroom Organization, Responsive Teaching, Proactive Management and Routines, and Cognitive Facilitation.

Table 6

Correlation Coefficients for the Ten CLASS Pre-K Dimensions

	PC	NC	TS	RSP	CD	QF	LM	BM	PD	ILF
PC	1									
NC	-.47	1								
TS	.64	-.43	1							
RSP	.63	-.45	.63	1						
CD	.35	-.18	.26	.30	1					
QF	.25	-.09	.26	.25	.74	1				
LM	.44	-.14	.49	.42	.61	.65	1			
BM	.61	-.43	.80	.67	.30	.24	.50	1		
PD	.65	-.43	.68	.72	.27	.32	.49	.78	1	
ILF	.52	-.37	.56	.63	.45	.48	.63	.73	.73	1

Note: A total of 86 out of the 145 Head Start Classrooms across 22 centers were included in this analysis. The dimension abbreviations are as follows: PC = Positive Climate, NC = Negative Climate, TS = Teacher Sensitivity, RSP = Regard for Child Perspectives, CD = Concept Development, QF = Quality of Feedback, LM = Language Modeling, BM = Behavior Management, PD = Productivity, ILF = Instructional Learning Formats.

Table 7

Factor Loadings, Variances, and Covariances for Current Study CFAs

	Three-Factor Model				**Bifactor Model**			
	Variances	**Factor Loadings**			**Variances**	**Factor Loadings**		
Observed Variables		ES	CO	IS		PMR	CF	RT
Positive Climate	.43	.76			.20	.49		.75
Negative Climate	.72	-.53			.71	-.20		-.50
Teacher Sensitivity	.31	.83			.31			.83
Regard for Child Perspectives	.35	.81			.40			.78
Behavior Management	.20		.90		.14	-.15		.92
Productivity	.22		.89		.25	.02		.86
Instructional Learning Formats	.35		.81		.39			.78
Concept Development	.32			.83	.33		.73	.37
Quality of Feedback	.27			.86	.19		.84	.33

Language Modeling	.41		.77	.38	.54	.58
Covariances						
ES		.96	.47			
CO			.52			

Note: Factor abbreviations are as follows: ES = Emotional Support, IS = Instructional Support, CO = Classroom Organization, RT = Responsive Teaching, PMR = Proactive Management and Routines, CF = Cognitive Facilitation.

Table 8

Fit Statistics for Current Study CFAs

Fit Indices	Three-Factor Model	Bifactor Model
χ^2	78.07	59.93
df	32	28
p	.00	.00
CFI	.92	.94
TLI	.89	.91
RMSEA	.13	.12
SRMR	.08	.06

Table 9

Correlation Matrix of Factors from the Three-Factor Model of CLASS Pre-K

	Emotional Support	Instructional Support	Classroom Organization
Emotional Support	1		
Instructional Support	.47	1	
Classroom Organization	.96	.52	1

Table 10

Multilevel Model Results for Total Protective Factors Change Scores

		Three-factor Model				Bifactor Model	
	Unconditional	1.1	1.2	1.3	1.4	2.1	2.2
Fixed Effects							
Intercept	1.20	**4.80***	**4.76***	**4.79***	**4.96***	**4.82***	**5.59***
Emotional Support		0.50			-3.73		
Instructional Support			0.30		-0.38		
Classroom Organization				0.59	4.53		
Responsive Teaching						0.48	0.50
Proactive Management and Routines							-0.90
Cognitive Facilitation							-0.05
Classroom Average Fall Total Protective Factors		**-0.65****	**-0.65****	**-0.65****	**-0.65****	**-0.65****	**-0.65****
Center ID 1		-5.87	-6.14	-5.89	-6.14	-5.86	-6.81
Center ID 2		-5.24	-4.69	-5.36	-5.97	-5.27	-6.07
Center ID 3		-4.42	-4.95	-4.45	-5.07	-4.56	-5.42
School ID 4		-4.48	-4.54	-4.47	-4.72	-4.50	-6.65
Center ID 5		-5.74	-6.21	-5.68	-6.05	-5.82	-6.75
Center ID 6		-0.36	-0.61	-0.33	-0.23	-0.35	-0.31
Center ID 7		-1.87	-1.54	-1.89	-2.60	-1.87	-2.85
Center ID 8		-3.76	-3.57	-3.55	-3.11	-3.72	-4.23
Center ID 9		-3.94	-3.90	-3.80	-3.34	-3.93	-4.25
Center ID 10		-3.30	-3.71	-3.17	-2.64	-3.32	-5.91
Center ID 11		0.41	0.27	0.40	0.30	0.40	-0.83
Center ID 12		-4.56	-4.51	-4.53	-5.07	-4.59	-5.82
Center ID 13		-4.70	-4.29	-4.75	-4.86	-4.67	-5.27
Center ID 14		-5.63	-5.30	-5.67	-6.16	-5.68	-6.96
Center ID 15		-4.64	-4.25	-4.66	-4.66	-4.65	-5.33
Center ID 16		-3.29	-3.15	-3.24	-3.41	-3.33	-4.36
Center ID 17		-6.24	-5.74	-6.35	-6.91	-6.26	-7.39
Center ID 18		-3.70	-3.86	-3.65	-3.36	-3.69	-3.73
Random Effects							
Classroom ID	**6.07****	**3.39****	**3.39****	**3.38****	**3.32****	**3.39****	**3.32****
Residual	**9.61****	**9.61****	**9.61****	**9.61****	**9.61****	**9.61****	**9.61****

Total R^2/Between R^2		0.19/0.61	0.19/0.61	0.19/0.61	0.20/0.61	0.19/0.61	0.20/0.61
ICC	0.38	0.26	0.26	0.26	0.26	0.26	0.26

Note: Center IDs were manually dummy coded and included as fixed effects to meet the input requirements of the r2mlm package in R, which does not internally handle categorical variables (Rights & Sterba, 2020). While multilevel modeling packages like lme4 typically convert categorical variables (e.g., Center ID) into dummy codes automatically (Bates et al., 2015), r2mlm requires these variables to be explicitly dummy coded prior to analysis (Rights & Sterba, 2020).

*$p < .05$

**$p < .01$

Table 11

Multilevel Model Results for Behavioral Concerns Change Scores

		Three-factor Model				Bifactor Model	
	Unconditional	3.1	3.2	3.3	3.4	4.1	4.2
Fixed Effects							
Intercept	1.08	-1.30	-1.07	-1.28	-1.12	-1.29	-1.67
Emotional Support		-0.03			3.90		
Instructional Support			-0.48		-0.63		
Classroom Organization				-0.14	-3.45		
Responsive Teaching						-0.07	0.17
Proactive Management and Routines							0.79
Cognitive Facilitation							-0.74
Classroom Average Fall Behavioral Concerns		**-0.63****	**-0.63****	**-0.63****	**-0.63****	**-0.63****	**-0.62****
Center ID 1		2.79	2.85	2.75	3.41	2.76	4.14
Center ID 2		3.40	3.16	3.50	2.95	3.44	3.11
Center ID 3		3.67	3.63	3.57	4.83	3.64	4.89
School ID 4		0.19	0.02	0.17	0.26	0.18	1.86
Center ID 5		4.04	3.61	3.87	4.42	3.97	4.87
Center ID 6		-1.85	-1.78	-1.90	-1.54	-1.87	-1.41
Center ID 7		0.58	-0.14	0.55	-0.08	0.57	0.33
Center ID 8		2.63	1.75	2.52	0.92	2.59	1.94
Center ID 9		5.16	4.75	5.08	4.21	5.13	5.00
Center ID 10		1.17	1.08	1.05	1.09	1.12	3.91
Center ID 11		0.35	0.39	0.33	0.66	0.34	1.63
Center ID 12		3.18	2.58	3.11	2.85	3.15	3.55
Center ID 13		3.19	3.11	3.27	2.68	3.22	3.22
Center ID 14		6.35	5.91	6.37	5.83	6.36	6.47
Center ID 15		1.74	1.54	1.80	1.01	1.77	1.62
Center ID 16		2.73	2.27	2.68	2.15	2.72	2.84
Center ID 17		1.75	1.54	1.84	1.36	1.79	1.73
Center ID 18		2.97	2.99	2.93	2.96	2.96	3.24
Random Effects							
Classroom ID	**4.83****	**2.90****	**2.87****	**2.89****	**2.78****	**2.90****	**2.77****
Residual	**9.72****	**9.72****	**9.72****	**9.72****	**9.72****	**9.72****	**9.72****

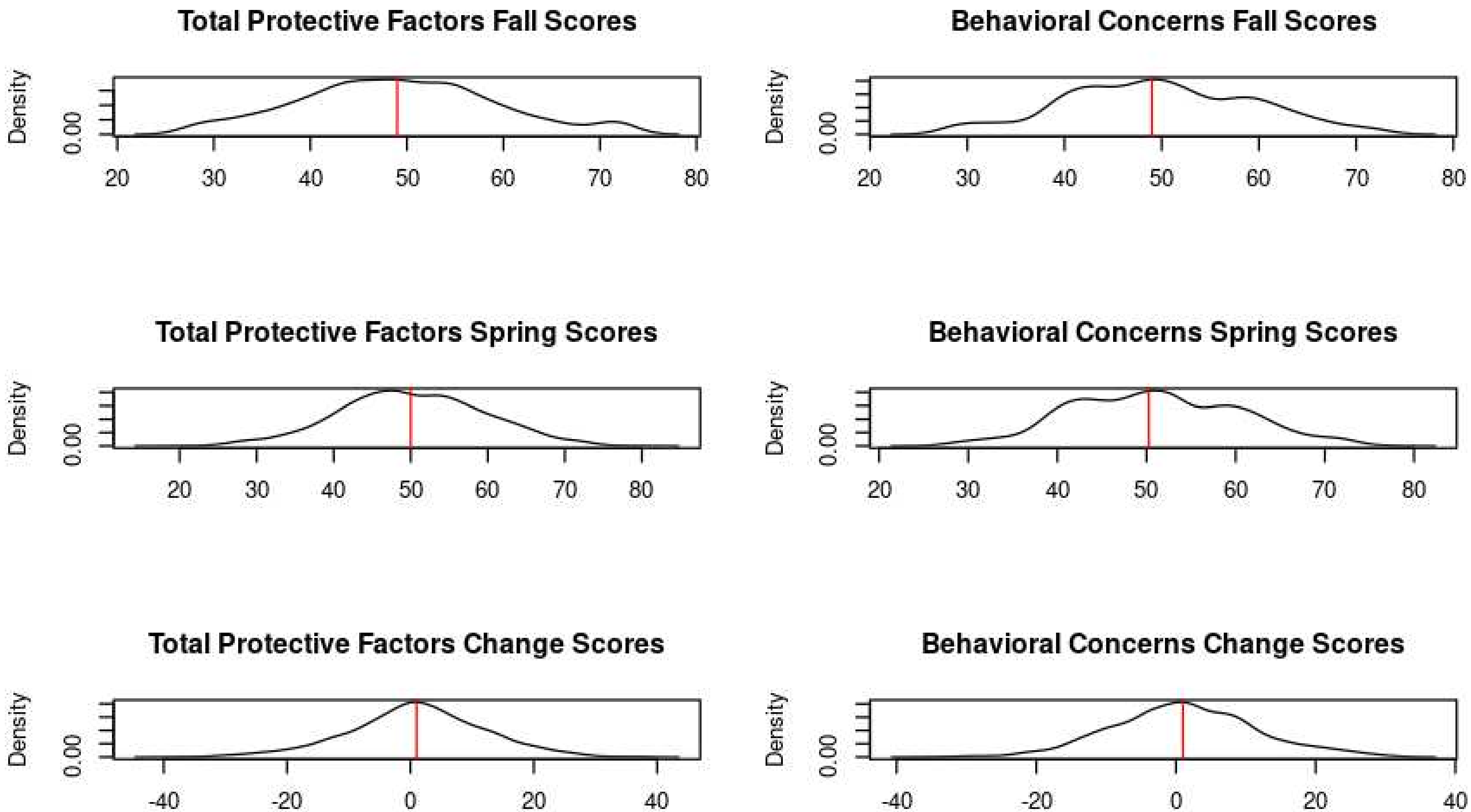
Total R^2/Between R^2		0.12/0.54	0.12/0.54	0.12/0.54	0.13/0.55	0.12/0.54	0.13/0.56
ICC	0.33	0.23	0.23	0.23	0.22	0.23	0.22

Note: Center IDs were manually dummy coded and included as fixed effects to meet the input requirements of the r2mlm package in R, which does not internally handle categorical variables (Rights & Sterba, 2020). While multilevel modeling packages like lme4 typically convert categorical variables (e.g., Center ID) into dummy codes automatically (Bates et al., 2015), r2mlm requires these variables to be explicitly dummy coded prior to analysis (Rights & Sterba, 2020).

*$p < .05$

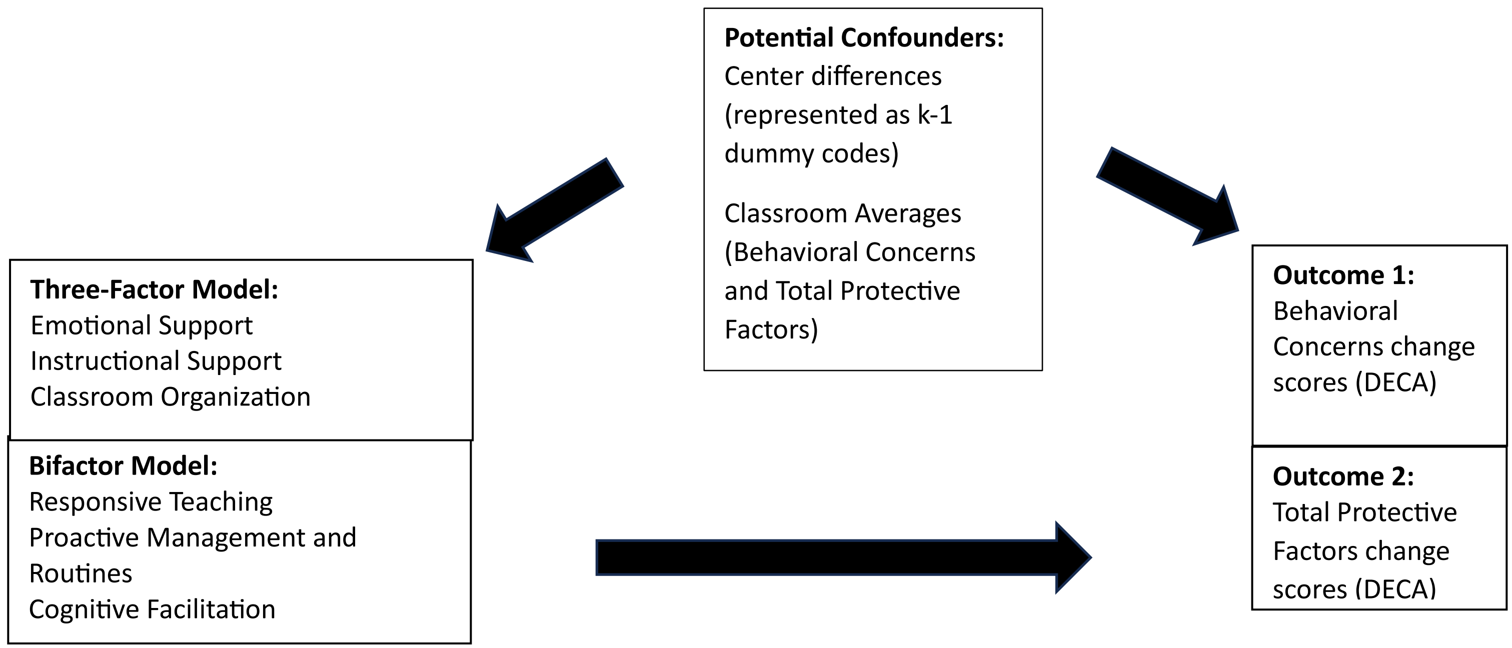
**$p <.01$

Figure 1



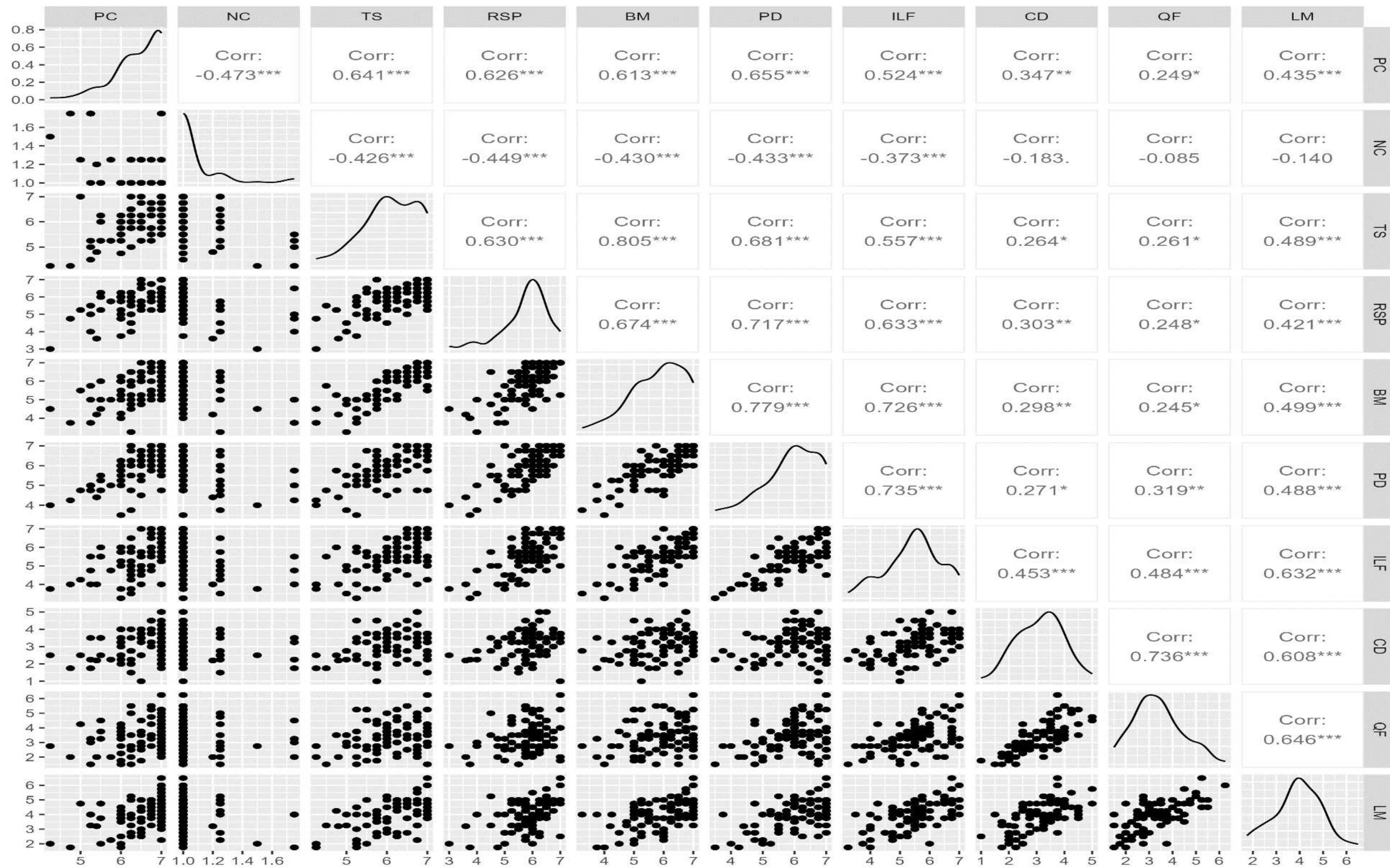
Density plots of Fall, Spring, and Change Scores for both outcomes of Total Protective Factors and Behavioral Concerns.
Note. The red vertical line represents the median score.

Figure 2



Relationship between CLASS Pre-K Three-Factor and Bifactor Models and Child Outcomes

Figure 3



Scatterplot Correlation Matrix of CLASS Pre-K Dimensions

Note: The dimension abbreviations are as follows: PC = Positive Climate, NC = Negative Climate, TS = Teacher Sensitivity, RSP = Regard for Child Perspectives, CD = Concept Development, QF = Quality of Feedback, LM = Language Modeling, BM = Behavior Management, PD = Productivity, ILF = Instructional Learning Formats.

REFERENCES

Abreu-Lima, I. M. P., Leal, T. B., Cadima, J., & Gamelas, A. M. (2013). Predicting child outcomes from preschool quality in Portugal. *European Journal of Psychology of Education, 28*(2), 399–420. http://www.jstor.org/stable/23421901

Administration for Children and Families. (2024a). *History of Head Start.* U.S. Department of Health & Human Services. https://www.acf.hhs.gov/ohs/about/history-head-start

Administration for Children and Families. (2024b). *Head Start Program Fact Sheet*. U.S. Department of Health & Human Services. https://www.acf.hhs.gov/ohs

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation*. Hillsdale, NJ: *Lawrence Erlbaum.*

Alavi, M., Visentin, D. C., Thapa, D. K., Hunt, G. E., Watson, R., & Cleary, M. (2020). Chi-square for model fit in confirmatory factor analysis. *Journal of advanced nursing, 76*(9), 2209-2211. https://doi.org/10.1111/jan.14399

Anders, Y., Rossbach, H.-G., Weinert, S., Ebert, S., Kuger, S., Lehrl, S., & von Maurice, J. (2012). Home and preschool learning environments and their relations to the development of early numeracy skills. *Early Childhood Research Quarterly, 27*(2), 231–244. https://doi.org/10.1016/j.ecresq.2011.08.003

Auger, A., Farkas, G., Burchinal, M. R., Duncan, G. J., & Vandell, D. L. (2014). Preschool center care quality effects on academic achievement: An instrumental variables analysis. *Developmental Psychology, 50*(12), 2559–2571. https://doi.org/10.1037/a0037995

Arend, M. G., Schäfer, T., Steinley, D., & Harlow, L. L. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, *24*(1), 1–19. https://doi.org/10.1037/met0000195

Acevedo-Garcia, D., & Hardy, E. (2016). Head Start: Taking on new challenges to address school readiness gaps. *Journal of Applied Research on Children: Informing Policy for Children at Risk,* 7(1), Article 11. https://www.diversitydatakids.org/sites/default/files/file/head-start-taking-on-new-challenges-to-address-school-readiness-gaps.pdf

Bagozzi, R. P. (1981). Evaluating structural equation models with unobservable variables and measurement error: A comment. *Journal of Marketing Research, 18* (3). https://doi.org/10.1177/002224378101800312

Bagozzi, R. P., & Phillips, L. W. (1982). Representing and testing organizational theories: A holistic construal. *Administrative Science Quarterly,* 459-489. https://doi.org/10.2307/2392322

Bailey, C. S., Ondrusek, A. R., Curby, T. W., & Denham, S. A. (2022). Teachers' consistency of emotional support moderates the association between young children's regulation capacities and their preschool adjustment. *Psychology in the Schools, 59*(6), 1051–1074. https://doi.org/10.1002/pits.22659

Baker, J. A., (2006). Contributions of teacher–child relationships to positive school adjustment during elementary school. *Journal of School Psychology, 44*(3), 211–229. https://doi.org/10.1016/j.jsp.2006.02.002

Barbu, O. C., Levine-Donnerstein, D., Marx, R. W., & Yaden, D. B. (2013). Reliability and validity of the Devereux Early Childhood Assessment (DECA) as a function of parent and teacher ratings. *Journal of Psychoeducational Assessment, 31*(5), 469–481. https://doi.org/10.1177/0734282912467758

Bassok, D., & Galdo, E. (2016). Inequality in preschool quality? Community-level disparities in access to high-quality learning environments. *Early Education and Development, 27*(1), 128–144. https://doi.org/10.1080/10409289.2015.1057463

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... & Bolker, M. B. (2015). Package 'lme4'. *Convergence, 12*(1), 2. http://dk.archive.ubuntu.com/pub/pub/cran/web/packages/lme4/lme4.pdf

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Loudermilk, B. L., & Kromrey, J. D. (2010). What is the impact of missing data on multilevel models? A demonstration using simulated data. *Journal of Experimental Education, 78*(3), 312–345. https://doi.org/10.1080/00220970903352760

Bender, S. L., Fedor, M. C., & Carlson, J. S. (2011). Examining Total Protective Factors and risk factors in urban and rural Head Start preschoolers. *Journal of Community Psychology, 39*(8), 908–921. https://doi.org/10.1002/jcop.20477

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin, 107*(2), 238. https://psycnet.apa.org/doi/10.1037/0033-2909.107.2.238

Berkowitz, R., Moore, H., Astor, R. A., & Benbenishty, R. (2017). A research synthesis of the associations between socioeconomic background, inequality, school climate, and academic achievement. *Review of Educational Research, 87*(2), 425–469. https://doi.org/10.3102/0034654316669821

Bettencourt, A. F., Gross, D., Ho, G., & Perrin, N. (2018). The costly consequences of not being socially and behaviorally ready to learn by kindergarten in Baltimore City. *Journal of Urban Health, 95*(1), 36–50. https://doi.org/10.1007/s11524-017-0214-6

Bierman, K. L., Nix, R. L., Heinrichs, B. S., Domitrovich, C. E., Gest, S. D., Welsh, J. A., & Gill, S. (2014). Effects of Head Start REDI on children's outcomes 1 year later in different kindergarten contexts. *Child Development*, *85*(1), 140–159. https://doi.org/10.1111/cdev.12117

Bihler, L. M., Agache, A., Kohl, K., Willard, J. A., & Leyendecker, B. (2018) Factor analysis of the Classroom Assessment Scoring System replicates the three domain structure and reveals no support for the bifactor model in German preschools. *Frontiers in Psychology, 9*(1232). https://doi.org/10.3389/fpsyg.2018.01232

Birch, S. H., & Ladd, G. W. (1997). The teacher-child relationship and children's early school adjustment. *Journal of School Psychology, 35*(1), 61–79. https://doi.org/10.1016/S0022-4405(96)00029-5

Birch, S. H., & Ladd, G. W. (1996*). Interpersonal relationships in the school environment and children's early school adjustment: The role of teachers and peers.* In J. Juvonen & K. R. Wentzel (Eds.), *Social motivation: Understanding children's school adjustment.* (p. 199–225). Cambridge University Press. https://doi.org/10.1017/CBO9780511571190.011

Blair, C., (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *The American Psychologist, 57*(2), 111–127. https://doi.org/10.1037/0003-066X.57.2.111

Bloom H. S. & Weiland C. (2015). *Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study.* Washington, DC: MDRC.

Bowlby, J. (1982). Attachment and loss: Retrospect and prospect. *American Journal of Orthopsychiatry, 52*(4), 664–678. https://doi.org/10.1111/j.1939-0025.1982.tb01456.x

Brinkman, T. M., Wigent, C. A., Tomac, R. A., Pham, A. V., & Carlson, J. S. (2007). Using the Devereux Early Childhood Assessment to identify behavioral risk and Total Protective Factors within a Head Start population. *Canadian Journal of School Psychology, 22*(2), 136–151. https://doi.org/10.1177/0829573507307612

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Browne, M.W. & Cudeck, R. (1992). *Alternative ways of assessing model fit.* In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (136-62), Newbury Park, CA: Sage

Buhs, E. S., Ladd, G. W., & Herald, S. L. (2006). Peer exclusion and victimization: Processes that mediate the relation between peer group rejection and children's classroom engagement and achievement? *Journal of Educational Psychology, 98*(1), 1–13. https://doi.org/10.1037/0022-0663.98.1.1

Buhs, E. S., & Ladd, G. W. (2001). Peer rejection as an antecedent of young children's school adjustment: An examination of mediating processes. *Developmental Psychology, 37*(4), 550–560. https://doi.org/10.1037/0012-1649.37.4.550

Bulotsky-Shearer, R. J., Fernandez, V. A., & Rainelli, S. (2013). The validity of the Devereux Early Childhood Assessment for culturally and linguistically diverse Head Start children. *Early Childhood Research Quarterly, 28*(4), 794–807. https://doi.org/10.1016/j.ecresq.2013.07.009

Bulotsky-Shearer, R. J., Wen, X., Faria, A. M., Hahs-Vaughn, D. L., & Korfmacher, J. (2012). National profiles of classroom quality and family involvement: A multilevel examination of proximal influences on Head Start children's school readiness. *Early Childhood Research Quarterly, 27*(4), 627-639. https://doi.org/10.1016/j.ecresq.2012.02.001

Burchinal M. R., Cryer D. (2003). Diversity, childcare quality, and developmental outcomes. *Early Childhood Research Quarterly, 18,* 401–426. https://doi.org/10.1016/j.ecresq.2003.09.003.

Burchinal, M. R., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher-child interactions and instruction. *Applied Developmental Science, 12*(3), 140–153. https://doi.org/10.1080/10888690802199418

Burchinal, M. R., Peisner-Feinberg, E., Pianta, R., & Howes, C. (2002). Development of academic skills from preschool through second grade: Family and classroom predictors of developmental trajectories. *Journal of School Psychology, 40*(5), 415–436. https://doi.org/10.1016/S0022-4405(02)00107-3

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly, 25(*2), 166–176. https://doi.org/10.1016/j.ecresq.2009.10.004

Bustamante, A. S., & Hindman, A. H. (2019). Classroom quality and academic school readiness outcomes in head start: The indirect effect of approaches to learning. *Early Education and Development, 30*(1), 19–35. https://doi.org /10.1080/10409289.2018.1540249

Cabell, S. Q., Justice, L. M., Piasta, S. B., Curenton, S. M., Wiggins, A., Turnbull, K. P., & Petscher, Y. (2011). The impact of teacher responsivity education on preschoolers' language and literacy skills. *American Journal of Speech-Language Pathology, 20*(4), 315–330. https://doi.org/10.1044/1058-0360(2011/10-0104)

Cadima, J., Aguiar, C., & Barata, M. C. (2018). Process quality in Portuguese preschool classrooms serving children at-risk of poverty and social exclusion and children with disabilities. *Early Childhood Research Quarterly, 45*, 93–105. https://doi.org/10.1016/j.ecresq.2018.06.007

Camilli, G., Vargas, S., Ryan, S., & Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record, 112*(3), 579-620. https://doi.org/10.1177/016146811011200303

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. American psychologist, 15(8), 546. https://psycnet.apa.org/doi/10.1037/h0048255

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological bulletin, 56(2), 81. https://psycnet.apa.org/doi/10.1037/h0046016

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: young adult outcomes from the Abecedarian Project. *Applied Developmental Science, 6*(1), 42–57. https://doi.org/10.1207/S1532480XADS0601_05

Carlson, K. D., & Herdman, A. O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods, 15*(1), 17-32. https://doi.org/10.1177/1094428110392383

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with accuracy. *Early Childhood Research Quarterly, 27*(3), 529-542. https://doi.org/10.1016/j.ecresq.2011.12.006

Cassidy, J., & Asher, S. R. (1992). Loneliness and peer relations in young children. *Child Development, 63*(2), 350–365. https://doi.org/10.1111/j.1467-8624.1992.tb01632.x

Cameron, C. E., Connor, C. M., Morrison, F. J., & Jewkes, A. M. (2008). Effects of classroom organization on letter–word reading in first grade. *Journal of School Psychology, 46*(2), 173–192. https://doi.org/10.1016/j.jsp.2007.03.002

Chain, J., Dopp, A., Smith, G., Woodland, S., & LeBuffe, P. (2010). *The Devereux Early Childhood Assessment literature review.* The Devereux Center for Resilient Children. https://centerforresilientchildren.org/wp-content/uploads/DECA-Literature-Review-6-28-10.pdf

Cheung, M. W. L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology, 5,* 1521. https://doi.org/10.3389/fpsyg.2014.01521

Choi, J. Y., Horm, D., Jeon, S., & Ryu, D. (2019). Do stability of care and teacher-child interaction quality predict child outcomes in Early Head Start? *Early Education and Development, 30*(3), 337–356. https://doi.org/10.1080/10409289.2018.1546096

Cloney, D., Nguyen, C., Adams, R. J., Tayler, C., Cleveland, G., & Thorpe, K. (2017). Psychometric Properties of the Classroom Assessment Scoring System (Pre-K): Implications for measuring interaction quality in diverse early childhood settings. *Journal of Applied Measurement*, *18*(3), 299–318.

Connor, C., Son, S.-H., Hindman, A. H., & Morrison, F. J. (2005). Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology, 43*(4), 343–375. https://doi.org/10.1016/j.jsp.2005.06.001

Crosnoe, R., Johnson, M. K., & Elder, G. H., Jr. (2004). Intergenerational bonding in school: the behavioral and contextual correlates of child-teacher relationships. *Sociology of Education, 77*(1), 60–81. https://doi-org.ezproxy2.library.colostate.edu/10.1177/003804070407700103

Curby, T. W., Rudasill, K. M., Edwards, T., & Pérez-Edgar, K. (2011). The role of classroom quality in ameliorating the academic and social risks associated with difficult temperament. *School Psychology Quarterly, 26*(2), 175–188. https://doi.org/10.1037/a0023042

Curby, T. W., LoCasale-Crouch, J., Konold, T. R., Pianta, R. C., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2009a). The relations of observed Pre-K classroom quality profiles to children's achievement and social competence. *Early Education and Development*, 20(2), 346–372. https://doi.org/10.1080/10409280802581284

Curby, T. W., Rimm-Kaufman, S. E., & Ponitz, C. C. (2009b). Teacher–child interactions and children's achievement trajectories across kindergarten and first grade. Journal of Educational Psychology, 101(4), 912–925. https://doi.org/10.1037/ a0016647

Davis, E. A., & Miyake, N. (2004). Explorations of scaffolding in complex classroom systems. *The Journal of the Learning Sciences, 13*(3), 265–272. https://doi.org/10.1207/s15327809jls1303_1

Department of Health and Human Services. (2020). *Preamble to the final rule on Designation Renewal System changes*. U.S. Administration for Children and Families. https://eclkc.ohs.acf.hhs.gov/sites/default/files/pdf/preamble-final-rule-designation-renewal-system-charges.pdf

Devereux Advanced Behavioral Health Center for Resilient Children. (2024). *Enhancing social and emotional development: Devereux Early Childhood Assessment Program.* Kaplan Early Learning Company. https://www.kaplanco.com/media/DECA_Manual.pdf

Devereux Advanced Behavioral Health Center for Resilient Children. (2017). *Understanding the alignment of the Head Start Early Learning Outcomes Framework: Ages birth to five & the Devereux Early Childhood Assessment (DECA) Program for infants, toddlers, and* *preschoolers. https://centerforresilientchildren.org/wp-content/uploads/DECA-HS-Learning-Outcomes-Framework-Alignment-Tool-1.pdf*

Devereux Center for Resilient Children. (n.d.). *DECA-P2 psychometric properties*. https://centerforresilientchildren.org/wp-content/uploads/2014/01/DECA-P2-Psychometric-Properties.pdf

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D. and Lautenbach, S. (2013), Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*: 27-46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Downer, J. T., López, M. L., Grimm, K., Hamagami, A., Pianta, R. C., & Howes, C. (2012). Observations of teacher child interactions in classrooms serving Latinos and dual language learners: Applicability of the Classroom Assessment Scoring System in diverse settings. *Early Childhood Research Quarterly, 27*, 21–32. https://doi.org/10.1016/j.ecresq.2011.07.005

Downer, J. T., Sabol, T. J., & Hamre, B. (2010). Teacher-child interactions in the classroom: toward a theory of within- and cross-domain links to children’s developmental outcomes. *Early Education and Development, 21*(5), 699–723. https://doi.org/10.1080/10409289.2010.497453

Duncan, G. J., & Magnuson, K. A. (2005). Can family socioeconomic resources account for racial and ethnic test score gaps? *The Future of Children, 15*(1), 35–54. https://doi.org/10.1353/foc.2005.0004

Dunn, K. J., & McCray, G. (2020). The place of the bifactor model in confirmatory factor analysis investigations into construct dimensionality in language testing. *Frontiers in psychology, 11*, 520666. https://doi.org/10.3389%2Ffpsyg.2020.01357

Educare. (2024). *About us.* https://www.educareschools.org/who-we-are/about-educare/

Emmer, E. D., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*(2), 103–112. https://doi.org/10.1207/S15326985EP3602_5

Enders, C. K., Mistler, S. A., Keller, B. T., & Harlow, L. L. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods, 21*(2), 222–240. https://doi.org/10.1037/met0000063

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Farrell, A. M., & Rudd, J. M. (2009). Factor analysis and discriminant validity: A brief review of some practical issues. *In Australia and New Zealand Marketing Academy Conference 2009.* https://www.researchgate.net/publication/40500590_Factor_analysis_and_discriminant_validity_A_brief_review_of_some_practical_issues

Ferretti, L. K., (2023). Environmental structure: Contributions of family routines and classroom organization to the social and academic development of low-income kindergarteners. *ProQuest Dissertations & Theses Global. (2779137768).*

https://www.proquest.com/dissertations-theses/environmental-structure-contributions-family/docview/2779137768/se-2.

Fiscella, K., & Kitzman, H. (2009). Disparities in academic achievement and health: The intersection of child education and health policy. *Pediatrics (Evanston)*, *123*(3), 1073–1080. https://doi.org/10.1542/peds.2008-0533

Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of head start. *The American Economic Review, 92*(4), 999–1012. https://doi.org/10.1257/00028280260344560

Gaskins, C.S., Herres, J., & Kobak, R. (2012). Classroom order and child learning in late elementary school: A multilevel transactional model of achievement trajectories. *Journal of Applied Developmental Psychology, 33*(5), 227–235. https://doi.org/10.1016/j.appdev.2012.06.002

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal Pre-K on cognitive development. *Developmental Psychology, 41*(6), 872–884. https://doi.org/10.1037/0012-1649.41.6.872

Gordon, R. A., & Peng, F. (2020). Evidence regarding the domains of the CLASS PreK in Head Start classrooms. *Early Childhood Research Quarterly*, *53*, 23-39. https://doi.org/10.1016/j.ecresq.2020.01.008

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493-498.

Greenwald, A., G. (1975). Consequences of prejudice against the null hypothesis. *Psychological bulletin, 82*(1), 1. https://psycnet.apa.org/doi/10.1037/h0076157

Gregory, A., & Weinstein, R. S. (2004). Connection and regulation at home and in school: Predicting growth in achievement for adolescents. *Journal of Adolescent Research, 19*(4), 405–427. https://doi.org/10.1177/0743558403258859

Gutman, L. M., & Sulzby, E. (2000). The role of autonomy-support versus control in the emergent writing behaviors of African-American kindergarten children. *Reading Research and Instruction, 39*(2), 170–183. https://doi.org /10.1080/19388070009558319

Hamre, B., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M. A., & Hamagami, A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. The Elementary School Journal, 113(4), 461–487. https://doi.org/10.1086/669616

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher-child interactions: associations with preschool children's development. *Child Development, 85*(3), 1257–1274. https://doi.org/10.1111/cdev.12184

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949–967. https://doi.org/10.1111/j.1467-8624.2005.00889.x

Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development, 72*(2), 625–638. https://doi.org/10.1111/1467-8624.00301

Hardy, M. A. (1993). *Regression with dummy variables* (93). India: SAGE Publications. https://ehost/ebookviewer/ebook?sid=862234cb-483c-40c0-9077-60ca58e9b479%40redis&vid=0&format=EB

Harel, O., & Zhou, X.-H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine, 26*(16), 3057–3077. https://doi.org/10.1002/sim.2787

Hatfield, B. E., Burchinal, M. R., Pianta, R. C., & Sideris, J. (2016). Thresholds in the association between quality of teacher–child interactions and preschool children's school readiness skills. *Early Childhood Research Quarterly, 36*, 561–571. https://doi.org/10.1016/j.ecresq.2015.09.005

Hatfield, B. E., Lower, J. K., Cassidy, D. J., & Faldowski, R. A. (2015). Inequities in access to quality early care and education: Associations with funding and community context. *Early Childhood Research Quarterly, 30*, 316-326. https://doi.org/10.1016/j.ecresq.2014.01.001

Head Start Program Facts: Fiscal Year 2021. (2021). https://eclkc.ohs.acf.hhs.gov/about-us/article/head-start-program-facts-fiscal-year-2021

Hindman, A. H., Miller, A. L., Froyen, L. C., & Skibbe, L. E. (2012). A portrait of family involvement during Head Start: Nature, extent, and predictors. *Early Childhood Research Quarterly, 27*(4), 654-667. https://doi.org/10.1016/j.ecresq.2011.11.002

Hindman, A. H., Skibbe, L. E., Miller, A., & Zimmerman, M. (2010). Ecological contexts and early learning: Contributions of child, family, and classroom factors during Head Start, to literacy and mathematics growth through first grade. *Early childhood research quarterly, 25*(2), 235-250. https://doi.org/10.1016/j.ecresq.2009.11.003

Hindman, A. H., & Wasik, B. A. (2015). Building vocabulary in two languages: An examination of Spanish-speaking dual language learners in Head Start. *Early Childhood Research Quarterly, 31*, 19-33.

Hindman, A. H., & Wasik, B. A. (2013). Vocabulary learning in Head Start: Nature and extent of classroom instruction and its contributions to children's learning. *Journal of school psychology, 51*(3), 387-405. https://doi.org/10.1016/j.jsp.2013.01.001

Hedges, L. V., & Rhoads, C. H. (2010). *Statistical power analysis in education research* (NCSER 2010-3006). U.S. Department of Education, Institute of Education Sciences, National Center for Special Education Research. https://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf

Hoff, H., (2003). The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*(5), 1368–1378. https://doi.org/10.1111/1467-8624.00612

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Hu, B. Y., Fan, X., Gu, C., & Yang, N. (2016). Applicability of the classroom assessment scoring system in Chinese preschools based on psychometric evidence. *Early Education and Development, 27*(5), 714–734. https://doi.org/10.1080/ 10409289.2016.1113069

Jalongo, M. R. (2021). The effects of COVID-19 on early childhood education and care: Research and resources for children, families, teachers, and teacher educators. *Early Childhood Education Journal, 49*(5), 763–774. https://doi.org/10.1007/s10643-021-01208-y

Joshi, P., Geronimo, K., & Acevedo-Garcia, D. (2016). Head Start since the War on Poverty: Taking on new challenges to address persistent school readiness gaps. *The Journal of Applied Research on Children*, *7*(1). https://doi.org/10.58464/2155-5834.1282

Karlsen, L., Rydland, V., Solheim Buøen, E., Vandell, D. L., & Lekhal, R. (2023). The factor structures of the Classroom Assessment Scoring System Pre-K and mature play observation tool in multi-ethnic Norwegian early childhood centers. *Journal of Early Childhood Research*, *22*(1). https://doi.org/10.1177/1476718X231195708

Kasim, R. M., & Raudenbush, S. W. (1998). Application of gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics*, *23*(2), 93–116. https://doi.org/10.2307/1165316

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. Journal of Experimental Social Psychology, 12(3), 247-252. https://doi.org/10.1016/0022-1031(76)90055-X

Keys, T. D., Farkas, G., Burchinal, M. R., Duncan, G. J., Vandell, D. L., Li, W., Ruzek, E. A., & Howes, C. (2013). Preschool center quality and school readiness: Quality effects and variation by demographic and child characteristics. Child Development, 84(4), 1171–1190. https://doi.org/10.1111/cdev.12048

Khalfaoui, A., García-Carrión, R., & Villardón-Gallego, L. (2021). A systematic review of the literature on aspects affecting positive classroom climate in multicultural early childhood education. *Early Childhood Education Journal, 49*(1), 71–81. https://doi.org/10.1007/s10643-020-01054-4

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2008). A multilevel study of predictors of child perceptions of school climate: The effect of classroom-level factors. *Journal of educational psychology, 100*(1), 96. https://psycnet.apa.org/doi/10.1037/0022-0663.100.1.96

Kuppens, T., & Yzerbyt, V. Y. (2014). Predicting variability: Using multilevel modelling to assess differences in variance. *European Journal of Social Psychology*, *44*(7), 691–700. https://doi.org/10.1002/ejsp.2028

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. https://doi.org/10.1111/1467-8624.00101

Ladd, H. F., & Goertz, M. E. (2015). Handbook of research in education finance and policy, 2nd edition. In *Routledge, Taylor & Francis Group*. Routledge, Taylor & Francis Group. https://doi.org/10.4324/9780203788684

Ladd, G. W., Kochenderfer, B. J., & Coleman, C. C. (1997). Classroom peer acceptance, friendship, and victimization: distinct relational systems that contribute uniquely to children's school adjustment? *Child Development, 68*(6), 1181–1197. https://doi.org/10.2307/1132300

Ladd, G. W., & Troop-Gordon, W. (2003). The role of chronic peer difficulties in the development of childrens psychological adjustment problems. *Child Development, 74*(5), 1344–1367. https://doi.org/10.1111/1467-8624.00611

LeBuffe, P. A., & Naglieri, J. A. (1999). The Devereux Early Childhood Assessment (DECA): A measure of within-child Total Protective Factors in preschool children. *NHSA Dialog: A Research-to-Practice Journal for the Early Intervention Field, 3*(1), 75-80. https://doi.org/10.1207/s19309325nhsa0301_10

LeBuffe, P. A., & Naglieri, J. A. (2012*). Devereux Early Childhood Assessment for Preschoolers, Second Edition: User's Guide and Technical Manual.* Lewisville, NC: Kaplan Early Learning Company.

LeBuffe, P. A., & Shapiro, V. B. (2004). *The Devereux Early Childhood Assessment for Preschoolers (DECA-P2): User's Guide*. Lewisville, NC: Kaplan Press.

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school.* Economic Policy Institute, 1660 L Street, NW, Suite 1200, Washington, DC 20036.

Leyva, D. Weiland, C., Barata, M., Yoshikawa, H., Snow, C., Treviño, E., & Rolla, A. (2015). Teacher-child interactions in Chile and their associations with prekindergarten outcomes. *Child Development, 86*(3), 781–799. https://doi.org/10.1111/cdev.12342

Li, H., Liu, J., & Hunter, C. V. (2020). A meta-analysis of the factor structure of the Classroom Assessment Scoring System (CLASS*). Journal of Experimental Education, 88*(2), 265–287. https://doi.org/10.1080/00220973.2018.1551184

Li, J., Yan, X. S., Chaudhary, D., Avula, V., Mudiganti, S., Husby, H., ... & Abedi, V. (2021). Imputation of missing values for electronic health record laboratory data*. NPJ Digital Medicine, 4*(1), 147. https://doi.org/10.1038/s41746-021-00518-0

Lien, M. T., & Carlson, J. S. (2009). Psychometric properties of the Devereux early childhood assessment in a head start sample. *Journal of Psychoeducational Assessment, 27*(5), 386-396. https://doi.org/10.1177/0734282909331754

LoCasale-Crouch, J., Konold, T., Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2007). Observed classroom quality profiles in state-funded pre-kindergarten programs and associations with teacher, program, and classroom characteristics. *Early Childhood Research Quarterly, 22*(1), 3–17. https://doi.org/10.1016/j.ecresq.2006.05.001

Lowenstein, A. E., Friedman-Krauss, A. H., Raver, C. C., Jones, S. M., & Pess, R. A. (2015). School climate, teacher-child closeness, and low-income children's academic skills in kindergarten. *Journal of Educational and Developmental Psychology, 5*(2), 89–108. https://doi.org/10.5539/jedp.v5n2p89

Luke, D. (2004). *Multilevel modeling.* Thousand Oaks, CA: Sage.

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). The persistence of preschool effects: Do subsequent classroom experiences matter? *Early Childhood Research Quarterly, 22*(1), 18–38. https://doi.org/10.1016/j.ecresq.2006.10.002

Magnuson, K. A., & Waldfogel, J. (2005). Early childhood care and education: Effects on ethnic and racial gaps in school readiness. *The Future of Children*, *15*(1), 169–196. https://doi.org/10.1353/foc.2005.0005

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

McCoach, D. B., Keith, T. Z., & Siegle, D. (2013). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling* (2nd ed.). Routledge.

McFarland, J., Hussar, B., Wang, X., Zhang, J., Wang, K., Rathbun, A., Barmer, A., Forrest, E., Mann, F., & Bullock, F. (2018). *The condition of education 2018 (NCES 2018-144).* U.S. Department of Education. Washington, DC: National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo. asp?pubid=2018144.

McKenny, A. F., Short, J. C., & Payne, G. T. (2013). Using computer-aided text analysis to elevate constructs: An illustration using psychological capital. *Organizational research methods, 16(*1), 152-184. https://doi.org/10.1177/1094428112459910

Mitchell, A. (2001). Prekindergarten programs in the states: Trends and issues. https://files.eric.ed.gov/fulltext/ED454948.pdf

Moen, A. L., Sheridan, S. M., Schumacher, R. E., & Cheng, K. C. (2019). Early childhood child–teacher relationships: What is the role of classroom climate for children who are disadvantaged? *Early Childhood Education Journal, 47*(3), 331–341. https://doi.org/10.1007/s10643-019-00931-x

National Head Start Association. (2023). 2023 *Head Start workforce brief: Behavioral and developmental challenges in Head Start classrooms.* https://nhsa.org/wp-content/uploads/2023/03/2023.02-Workforce-Brief.pdf

Nezlek, J. B. (2012). Multilevel modeling for psychologists. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 3. Data analysis and research publication* (219–241). American Psychological Association. https://doi.org/10.1037/13621-011

Ng, E. L., Bull, R., Bautista, A., & Poon, K. (2021). A bifactor model of the Classroom Assessment Scoring System in preschool and early intervention classrooms in Singapore. *International Journal of Early Childhood*, *53*(2), 197-218. https://doi.org/10.1007/s13158-021-00292-w

Nores, M., & Barnett, W. S. (2010). Benefits of early childhood interventions across the world: (Under) Investing in the very young. *Economics of Education Review, 29*(2), 271–282. https://doi.org/10.1016/j.econedurev.2009.09.001

Office of Head Start (2024b). *Dual Language Learners in Head Start Programs.* Head Start Early Childhood Knowledge and Learning Center. https://eclkc.ohs.acf.hhs.gov/culture-language/article/visual-guide-dual-language-learners-head-start-programs

Office of Head Start (2024a). *Head Start program performance standards: Basis for determining whether a Head Start agency will be subject to an open competition.* Head Start Early Childhood Learning & Knowledge Center. https://eclkc.ohs.acf.hhs.gov/policy/45-cfr-chap-xiii/1304-11-basis-determining-whether-head-start-agency-will-be-subject-open-competition

Office of Head Start. (n.d.). *Head Start CLASS® and quality improvement practices*. U.S. Department of Health & Human Services. Retrieved from https://headstart.gov/teaching-practices/article/head-start-class-quality-improvement

Office of Head Start. (2023). *Use of Classroom Assessment Scoring System (CLASS®) in Head Start.* U.S. Department of Health and Human Services. https://eclkc.ohs.acf.hhs.gov/designation-renewal-system/article/use-classroom-assessment-scoring-system-class-head-start

Office of Head Start. (2015). *Head Start Early Learning Outcomes Framework Implementation Toolkit*. U.S. Department of Health and Human Services. https://eclkc.ohs.acf.hhs.gov/school-readiness/article/head-start-early-learning-outcomes-framework-implementation-toolkit

Ofsted. (2024). *School starters born during pandemic lack communication skills*. The Guardian. https://www.theguardian.com/education/2024/oct/08/school-starters-pandemic-lack-communication-skills-ofsted

Ogg, J.L., Brinkman, T. M., Dedrick, R. F., & Carlson, J. S. (2010). Factor structure and invariance across gender of the Devereux Early Childhood Assessment Protective Factor Scale. *School Psychology Quarterly, 25*(2), 107–118. https://doi.org/10.1037/a0020251

Pakarinen, E., Lerkkanen, M.-K., Poikkeus, A.-M., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., & Nurmi, J.-E. (2010). A validation of the Classroom Assessment Scoring System in Finnish kindergartens. *Early Education and Development, 21*(1), 95–124. https://doi.org/10.1080/10409280902858764

Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist, 36*(2), 89–101. https://doi.org/10.1207/S15326985EP3602_4

Pastori, G., & Pagani, V. (2017). Is validation always valid? Cross-cultural complexities of standard-based instruments migrating out of their context. *European Early Childhood Education Research Journal*, *25*(5), 682–697. https://doi.org/10.1080/1350293X.2017.1356545

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., & Yazejian, N. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development, 72*(5), 1534–1553. https://doi.org/10.1111/1467-8624.00364

Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. S. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the Classroom Assessment Scoring System) in early childhood education and care settings and child outcomes. *PloS One*, *11*(12), e0167660-. https://doi.org/10.1371/journal.pone.0167660

Pianta, R. C., Hamre, B. K., & Nguyen, T. (2020). Measuring and improving quality in early care and education. *Early Childhood Research Quarterly, 51*, 285–287. https://doi.org/10.1016/j.ecresq.2019.10.013

Pianta, R. C., Downer, J., & Hamre, B. (2016). Quality in early education classrooms: Definitions, gaps, and systems. *The Future of Children, 26*(2), 119–137. https://doi.org/10.1353/foc.2016.0015

Pianta R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008a). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal, 45*(2), 364–397. https://doi.org/10.3102/0002831207308230

Pianta, R.C., La Paro, K.M., Hamre, B.K. (2008b). *Classroom Assessment Scoring System: Manual Pre-K.* Paul H Brookes Publishing Co., Inc.

Pianta, R. C., Belsky, J., Houts, R., & Morrison, F. (2007a). Opportunities to learn in America's elementary classrooms. *Science (American Association for the Advancement of Science), 315*(5820), 1795–1796. https://doi.org/10.1126/science.1139719

Pianta, R. C., Cox, M. J., & Snow, K. L. (Eds.). (2007b). School readiness and the transition to kindergarten in the era of accountability (3-10). Baltimore, MD: Paul H. *Brookes Publishing.*

Pianta, R. C., Howes, C., Burchinal, M., 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. https://doi.org/10.1207/s1532480xads0903_2

Pianta, R. C., & Stuhlman, M. W. (2004a). Teacher-child relationships and children's success in the first years of school. *School Psychology Review, 33*(3), 444–458. https://doi.org/10.1080/02796015.2004.12086261

Pianta, R.C., la Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *The Elementary School Journal, 102*(3), 225–238. https://doi.org/10.1086/499701

Pianta, R. C. (1999). Enhancing relationships between children and teachers. *American Psychological Association.* https://doi.org/10.1037/10314-000

Pianta, R. C., & Stuhlman, M. W. (2004b). Teacher-child relationships and children's success in the first years of school. *School Psychology Review, 33*(3), 444–458. https://doi.org/10.1080/02796015.2004.12086261

Pinto, A. I., Pessanha, M., & Aguiar, C. (2013). Effects of home environment and center-based child care quality on children's language, communication, and literacy outcomes. *Early*

Childhood Research Quarterly, 28(1), 94–101. https://doi.org/10.1016/j.ecresq.2012.07.001

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., ... & Spier, E. (2010). *Head Start Impact Study. Final Report*. Administration for Children & Families. https://files.eric.ed.gov/fulltext/ED507845.pdf

Qi, C. H., Zieher, A., Lee Van Horn, M., Bulotsky-Shearer, R., & Carta, J. (2020). Language skills, behaviour problems, and classroom emotional support among preschool children from low-income families. *Early Child Development and Care, 190*(14), 2278–2290. https://doi.org/10.1080/03004430.2019.1570504

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raver, C. C., (2004). Placing Emotional Self-Regulation in Sociocultural and Socioeconomic Contexts. *Child Development, 75*(2), 346–353. https://doi.org/10.1111/j.1467-8624.2004.00676.x

Reichardt, C. S., & Coleman, S. C. (1995). The criteria for convergent and discriminant validity in a multitrait-multimethod matrix. *Multivariate Behavioral Research, 30(*4), 513-538. https://doi.org/10.1207/s15327906mbr3004_3

Reise, S. P., Bonifay, W., & Haviland, M. G. (2018). Bifactor modelling and the evaluation of scale scores. *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, 675-707. https://doi.org/10.1002/9781118489772.ch22

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6), 544-559. https://doi.org/10.1080/00223891.2010.496477

Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods, 25*(1), 6-14. https://doi.org/10.1177/1094428120968614

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods, 24*(3), 309–338. https://doi.org/10.1037/met0000184

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology, 45(*4), 958–972. https://doi.org/10.1037/a0015861

Rimm-Kaufman, S. E., La Paro, K. M., Downer, J. T., & Pianta, R. C. (2005). The contribution of classroom setting and quality of instruction to children's behavior in kindergarten classrooms. *The Elementary School Journal, 105*(4), 377–394. https://doi.org/10.1086/429948

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. https://doi.org/10.1016/S0193-3973(02)00128-4

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin, 86*(3), 638. https://psycnet.apa.org/doi/10.1037/0033-2909.86.3.638

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of statistical software, 48*, 1-36.

Rubin, D. B. (2018). Multiple imputation. *In Flexible Imputation of Missing Data, Second Edition* (29-62). Chapman and Hall/CRC. https://doi.org/10.1201/9780429492259

Rucinski, C.L., Brown, J. L., & Downer, J. T. (2018). Teacher-child relationships, classroom climate, and children's social-emotional and academic development. *Journal of Educational Psychology, 110*(7), 992–1004. https://doi.org/10.1037/edu0000240

Russell, J. A. (1978). Evidence of convergent validity on the dimensions of affect. *Journal of Personality and Social Psychology, 36*(10), 1152. https://doi.org/10.1037/0022-3514.36.10.1152

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Sabol, T. J., Hong, S. L. S., Pianta, R. C., & Burchinal, M. R. (2013). Can rating pre-k programs predict children's learning? *Science (American Association for the Advancement of Science)*, *341*(6148), 845–846. https://doi.org/10.1126/science.1233517Sabol, T. J., Ross, E. C., & Frost, A. (2020). Are all head start classrooms created equal? Variation in classroom quality within Head Start centers and implications for accountability systems. *American Educational Research Journal, 57*(2), 504–534. https://doi-org.ezproxy2.library.colostate.edu/10.3102/0002831219858920

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

Sandilos, L. E., Cycyk, L. M., Hammer, C. S., Sawyer, B. E., & Blair, C. (2016). Investigating the factor structure of the Classroom Assessment Scoring System (CLASS) in Head Start programs. *Early Education and Development, 27*(6), 775–792. https://doi.org/10.1080/10409289.2016.1141612

Schmitt N., Stults D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*(1), 1–22. https://doi.org/10.1177/014662168601000101

Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of clinical psychology, 62*(1), 123-146. https://doi.org/10.1002/jclp.20213

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organizational Research Methods, 19*(1), 80-110. https://doi.org/10.1177/1094428115598239

Shaw, M., Rights, J. D., Sterba, S. S., & Flake, J. K. (2022). r2mlm: An R package calculating R-squared measures for multilevel models. *Behavior Research Methods, 55*(4), 1942-1964. https://doi.org/10.3758/s13428-022-01841-4

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. https://doi.org/10.1016/j.jsp.2004.11.003

Skibbe, L., Behnke, M., & Justice, L. M. (2004). Parental scaffolding of children's phonological awareness skills: interactions between mothers and their preschoolers with language difficulties. *Communication Disorders Quarterly, 25*(4), 189–203. https://doi.org/10.1177/15257401040250040401

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and child engagement across the school year. *Journal of Educational Psychology, 85*(4), 571–581. https://doi.org/10.1037/0022-0663.85.4.571

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Stites, M. L., & Brown, E. T. (2021). Observing mathematical learning experiences in preschool. *Early Child Development and Care, 191*(1), 68–82. https://doi.org/10.1080/03004430.2019.1601089

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1-25.

Vorkapić, S. T. & LoCasale-Crouch, J. (Eds.). (2021). Supporting children's well-being during early childhood transition to school. *IGI Global.* https://doi.org/10.4018/978-1-7998-4435-8

Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal, 104*(1), 3–28. https://doi.org/10.1086/499740

Temple, J. A., Ou, S.-R., & Reynolds, A. J. (2022). Closing achievement gaps through preschool-to-third-grade programs. *Frontiers in Education (Lausanne)*, *7*. https://doi.org/10.3389/feduc.2022.871973

Tingstrom, D. H., Sterling-Turner, H. E., & Wilczynski, S. M. (2006). The good behavior game: 1969-2002. *Behavior Modification, 30*(2), 225–253. https://doi.org/10.1177/0145445503261165

Ulferts, H., Wolf, K. M., & Anders, Y. (2019). Impact of process quality in early childhood education and care on academic outcomes: longitudinal meta-analysis. *Child Development, 90*(5), 1474–1489. https://doi.org/10.1111/cdev.13296

U.S. Department of Health and Human Services (2022). *Home Language Support.* Administration for Children and Families. https://headstart.gov/culture-language/article/home-language-support?redirect=eclkc#:~:text=Children%20who%20speak%20home%20languages,to%20support%20home%20language%20development.

Valentino, R. (2018). Will public Pre-K really close achievement gaps? Gaps in prekindergarten quality between children and across states. *American Educational Research Journal, 55*(1), 79–116. https://doi.org/10.3102/0002831217732000

Valeski, T. N., & Stipek, D. J. (2001). Young children's feelings about school. *Child Development, 72*(4), 1198–1213. https://doi.org/10.1111/1467-8624.00342 van Buuren, S. (2018). *Flexible Imputation of Missing Data,* Second Edition. United Kingdom: CRC Press.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software, 45*, 1-67. https://doi.org/10.18637/jss.v045.i03

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology, 34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

Wang, L. Degol, J., Amemiya, J., Parr, A., & Guo, J. (2020). Classroom climate and children's academic and psychological wellbeing: A systematic review and meta-analysis. *Developmental Review, 57*. https://doi.org/10.1016/j.dr.2020.100912

Wang, Y., & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods, 20*(1), 63–83. https://doi.org/10.1037/met0000030

Wang, S., Wang, X., Liang, L., & Xie, H. (2024). Teacher-child interaction domains measured by the CLASS and children's pre-literacy skills: A systematic review and meta-analysis. *Early Education and Development,* 1–23. https://doi.org/10.1080/10409289.2022.2159259.

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. https://doi.org/10.1111/cdev.12099

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science, 24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Wentzel, K. R., (2003). Are effective teachers like good parents? Teaching styles and child adjustment in early adolescence. *Child Development, 73(*1), 287–301. https://doi.org/10.1111/1467-8624.00406

Werts C. E., Linn R. L. (1970). Path analysis: Psychological examples. *Psychological Bulletin, 74*(3), 193–212. https://doi.org/10.1037/h0029778

Wharton-McDonald, A., Pressley, M., & Hampston, J. M. (1998). Literacy instruction in nine first-grade classrooms: Teacher characteristics and child achievement. *The Elementary School Journal, 99*(2), 101–128. https://doi.org/10.1086/461918

Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior research methods, 55*(2), 788–806. https://doi.org/10.3758/s13428-022-01849-w

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual review of sociology, 25*(1), 659-706. https://doi.org/10.1146/annurev.soc.25.1.659

Wolf, G., & Cartwright, B. (1974). Rules for coding dummy variables in multiple regression. *Psychological Bulletin, 81*(3), 173–179. https://doi.org/10.1037/h0035848

Zaslow, M., Tout, K., Halle, T., Whittaker, J. V., & Lavelle, B. (2010). *Toward the identification of features of effective professional development for early childhood educators: Literature review*. Office of Planning, Research and Evaluation, Administration for

Children and Families, U.S. Department of Health and Human Services. https://www.acf.hhs.gov/sites/default/files/documents/opre/prof_dev_lit_review.pdf

Zatto, B. R. L., & Hoglund, W. L. G. (2019). Children's internalizing problems and teacher–child relationship quality across preschool. *Early Childhood Research Quarterly, 49*, 28–39. https://doi.org/10.1016/j.ecresq.2019.05.007

Zhai, F., Brooks-Gunn, J., & Waldfogel, J. (2011). Head Start and urban children's school readiness: A birth cohort study in 18 cities. *Developmental Psychology*, *47*(1), 134–152. https://doi.org/10.1037/a0020784

Zhai, F., Brooks-Gunn, J., & Waldfogel, J. (2014). Head Start's impact is contingent on alternative type of care in comparison group. *Developmental Psychology, 50*(12), 2572–2586. https://doi.org/10.1037/a0038205

Zheng, Q., Zheng, X., Liu, N., Fen, W. F., & Zhao Y. (2023). A meta-analysis of teacher–child interaction and early child outcomes from the perspective of CLASS. *Early Child Development and Care, 193*, 15-16, 1620-1651. https://doi.org/10.1080/03004430.2023.2273788

APPENDIX A

Follow-up Analyses

A series of scatterplots and boxplots were created to investigate the relationships between variables in this analysis and visualize classroom level variability in child outcome scores.

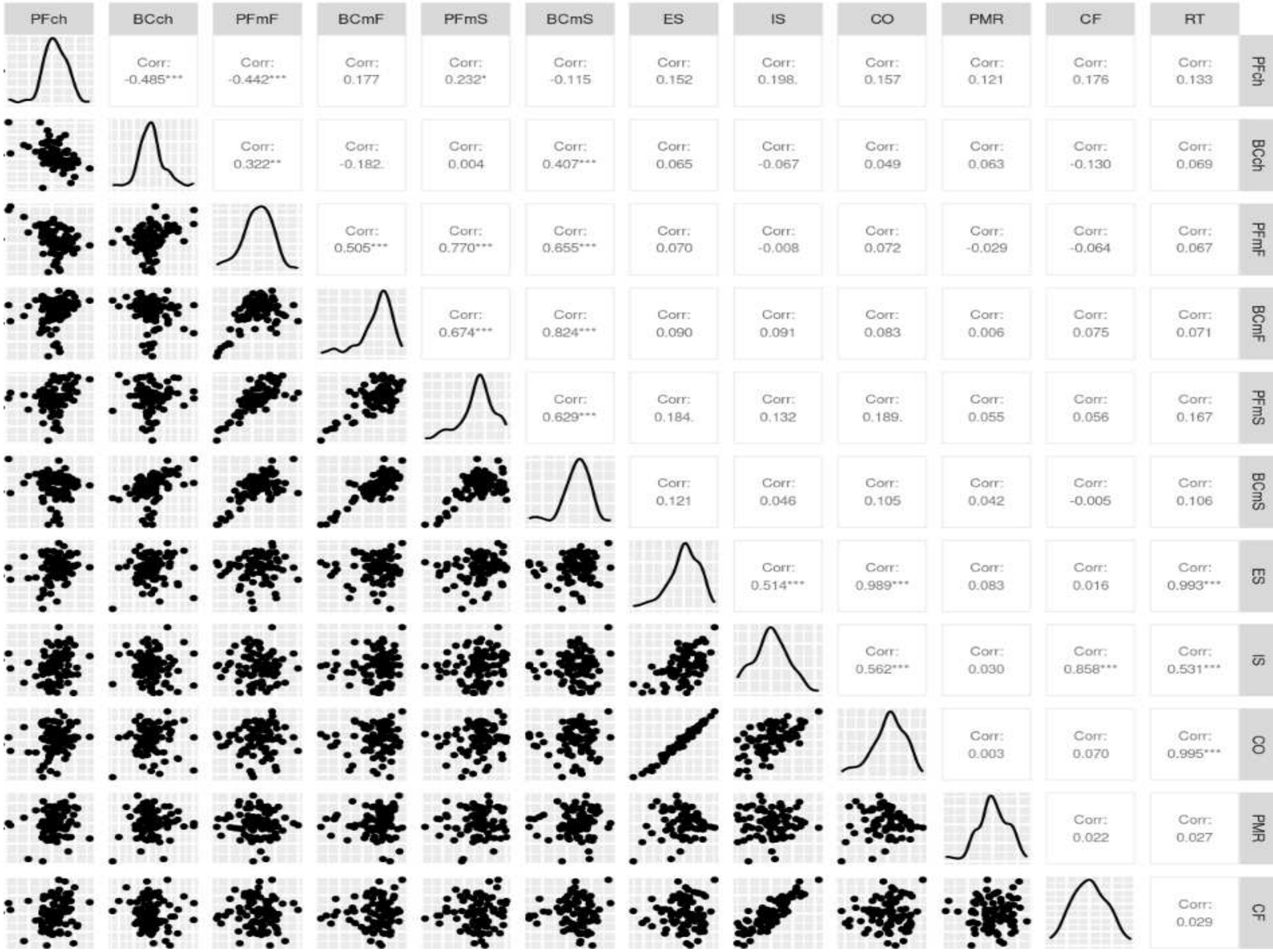
Scatterplots

Figure 4 is a scatterplot matrix of CLASS Pre-K factors and the classroom average pre, post, and change outcome scores. The scatterplot matrix reveals five high correlations; Total Protective Factors Fall classroom average scores and Total Protective Factors classroom average Spring scores (r = .88), Behavioral Concerns classroom average Fall scores and Behavioral Concerns classroom average Spring scores (r = .89), Emotional Support and Classroom Organization (r = .99), Emotional Support and Responsive Teaching (r = .99), and Classroom Organization and Responsive Teaching (r = .99).

It was expected for the Fall and Spring outcome scores to be correlated. Emotional Support and Classroom Organization are highly correlated and may be redundant when included in models together. According to the manual for CLASS Pre-K, these should be distinct factors measuring separate aspects of classroom quality (Pianta et al., 2008b). However, such a large correlation suggests they are measuring almost identical constructs.

Furthermore, the large correlations between Emotional Support, Classroom Organization, and Responsive Teaching suggest they are also measuring remarkably similar constructs. While some degree of correlation is expected since Responsive Teaching is used as a global measure of classroom quality and includes all dimensions, correlations of r = .99 means there are almost no differences between these factors. The exceedingly high correlations among predictors of CLASS Pre-K suggest further investigation of the dimensions and corresponding factors.

Figure 4

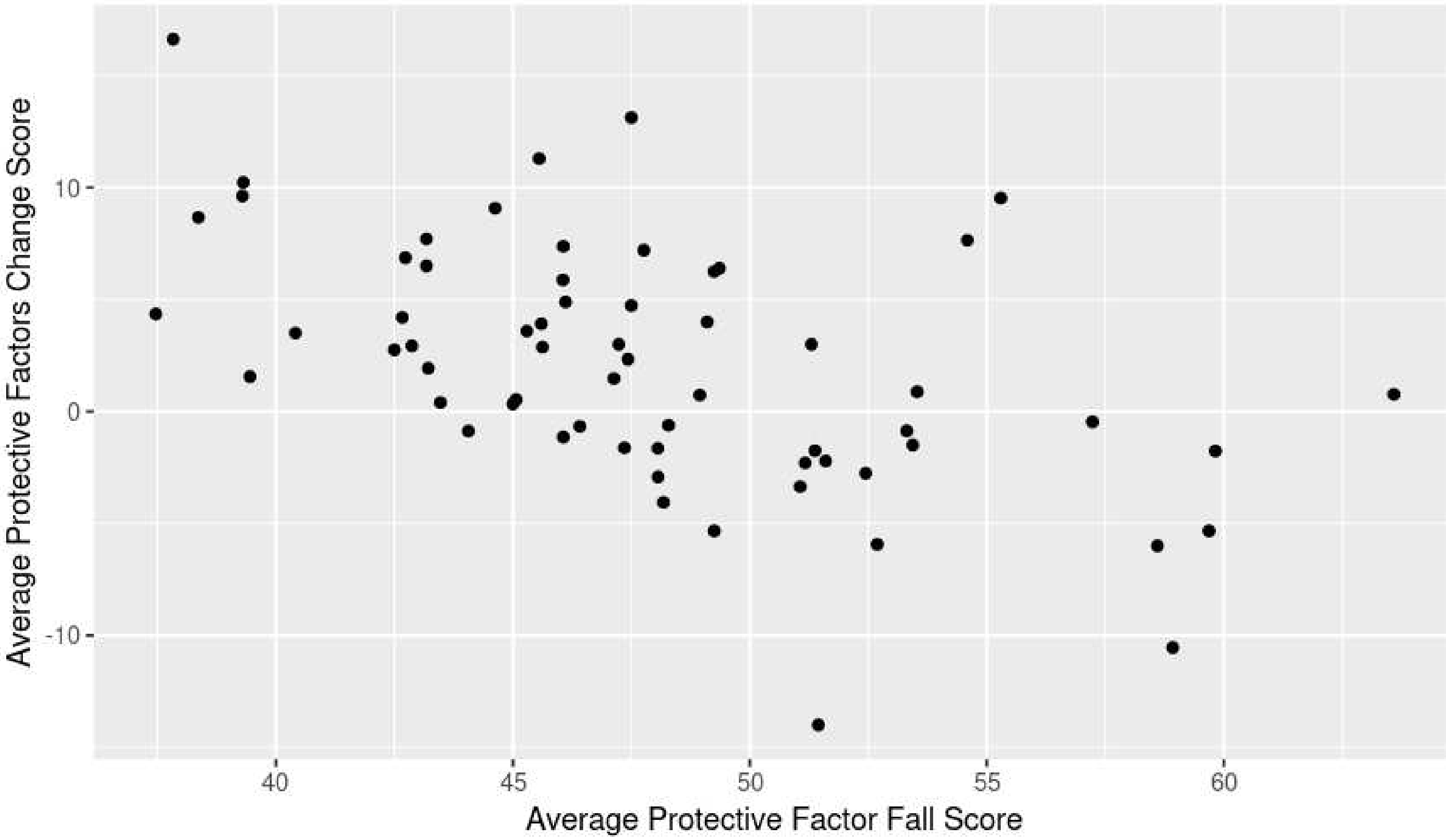


Scatterplot Correlation Matrix of CLASS Pre-K Factors and Classroom Average Pre, Post, and Change Outcome Scores

Note: The variable name abbreviations are as follows: PFch = classroom average of Total Protective Factors change scores, BCch = classroom average of Behavioral Concerns change scores, PFmF = classroom average of Total Protective Factors Fall scores, BCmF = classroom average of Behavioral Concerns Fall scores, PFmS = classroom average Total Protective Factors Spring scores, BCmS = classroom average Behavioral Concerns Spring scores, ES = Emotional Support classroom level factor scores, IS = Instructional Support classroom level factor scores, CO = Classroom Organization classroom level factor scores, PMR = Proactive Management and Routines classroom level factor scores, CF = Cognitive Facilitation classroom level factor scores, RT = Responsive Teaching classroom level factor scores.

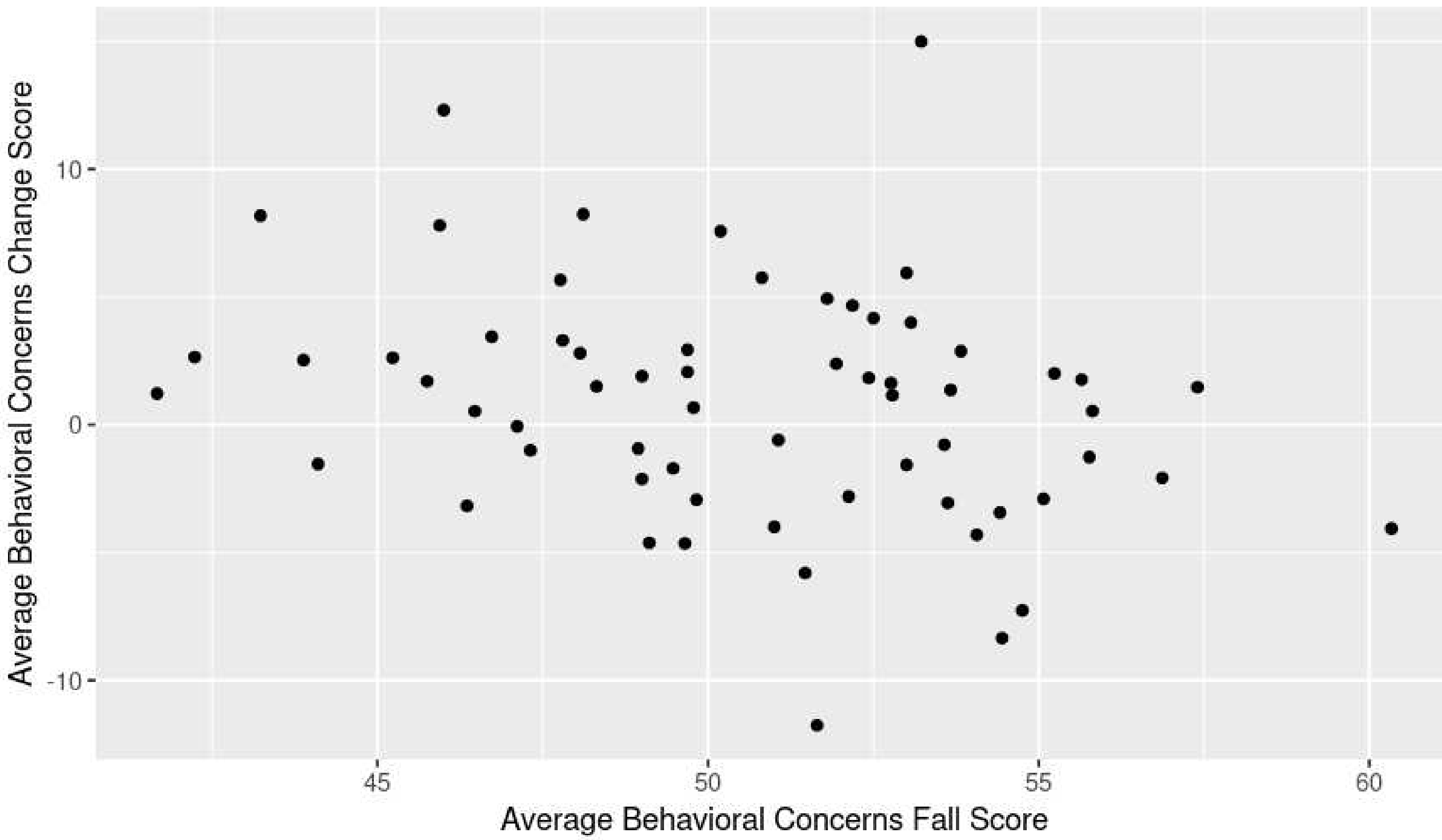
Figure 5 and Figure 6 were created using unimputed data, with one scatterplot each depicting Total Protective Factors and Behavioral Concerns, with the classroom level average DECA scores in Fall on the x-axis and the average change score for each classroom on the y-axis. These plots were used to investigate the unexpected negative effect of average classroom outcome scores. Figure 4 shows a negative relationship where classrooms that start out with lower Total Protective Factors scores in Fall tend to produce larger change scores, perhaps because they have more room for improvement. Figure 5 does not show a clear relationship between classroom average Behavioral Concerns Fall scores and change scores.

Figure 5



Scatterplot of Classroom Average Total Protective Factors Fall and Change Scores

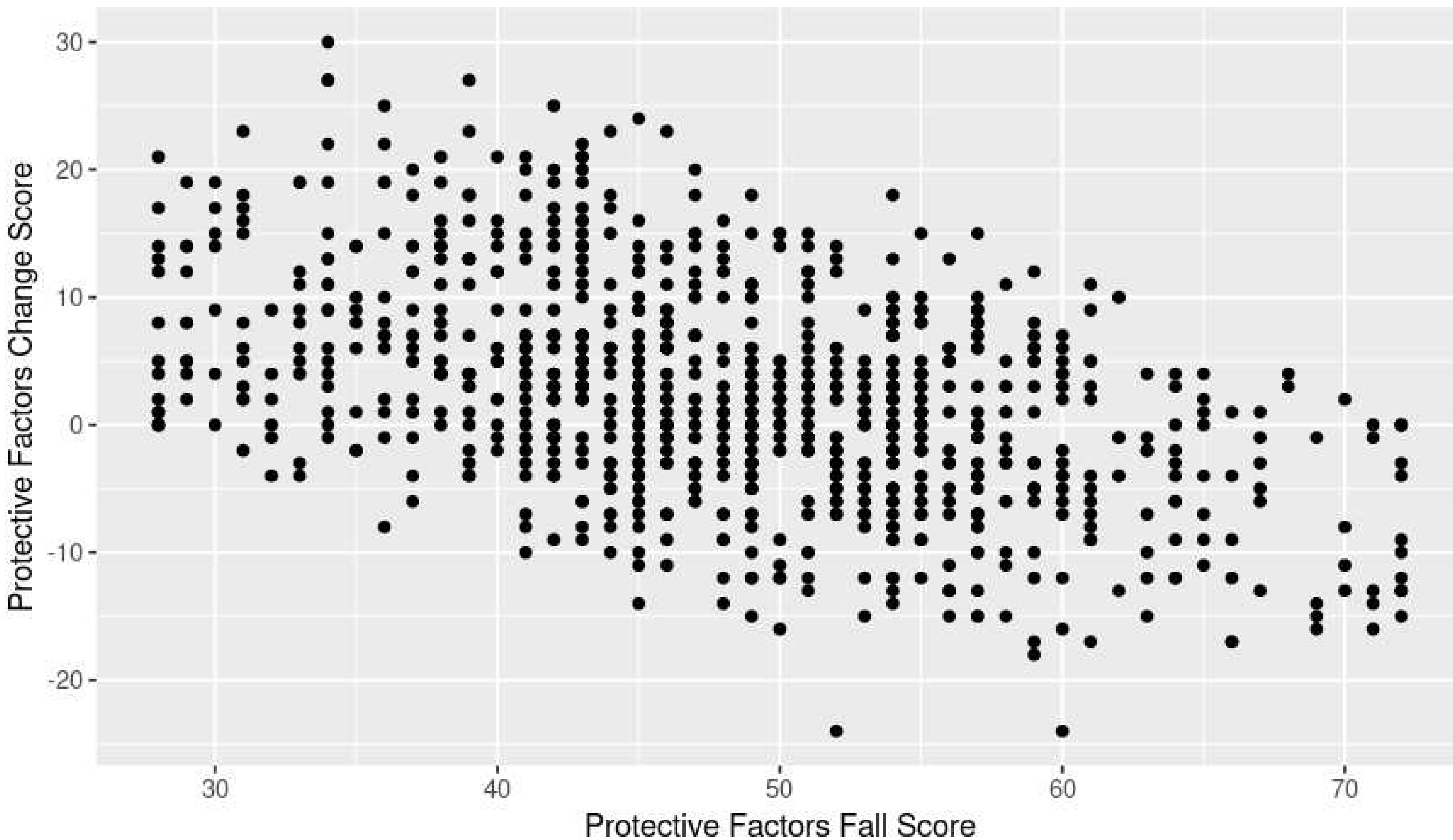
Figure 6



Scatterplot of Classroom Average Behavioral Concerns Fall and Change Scores

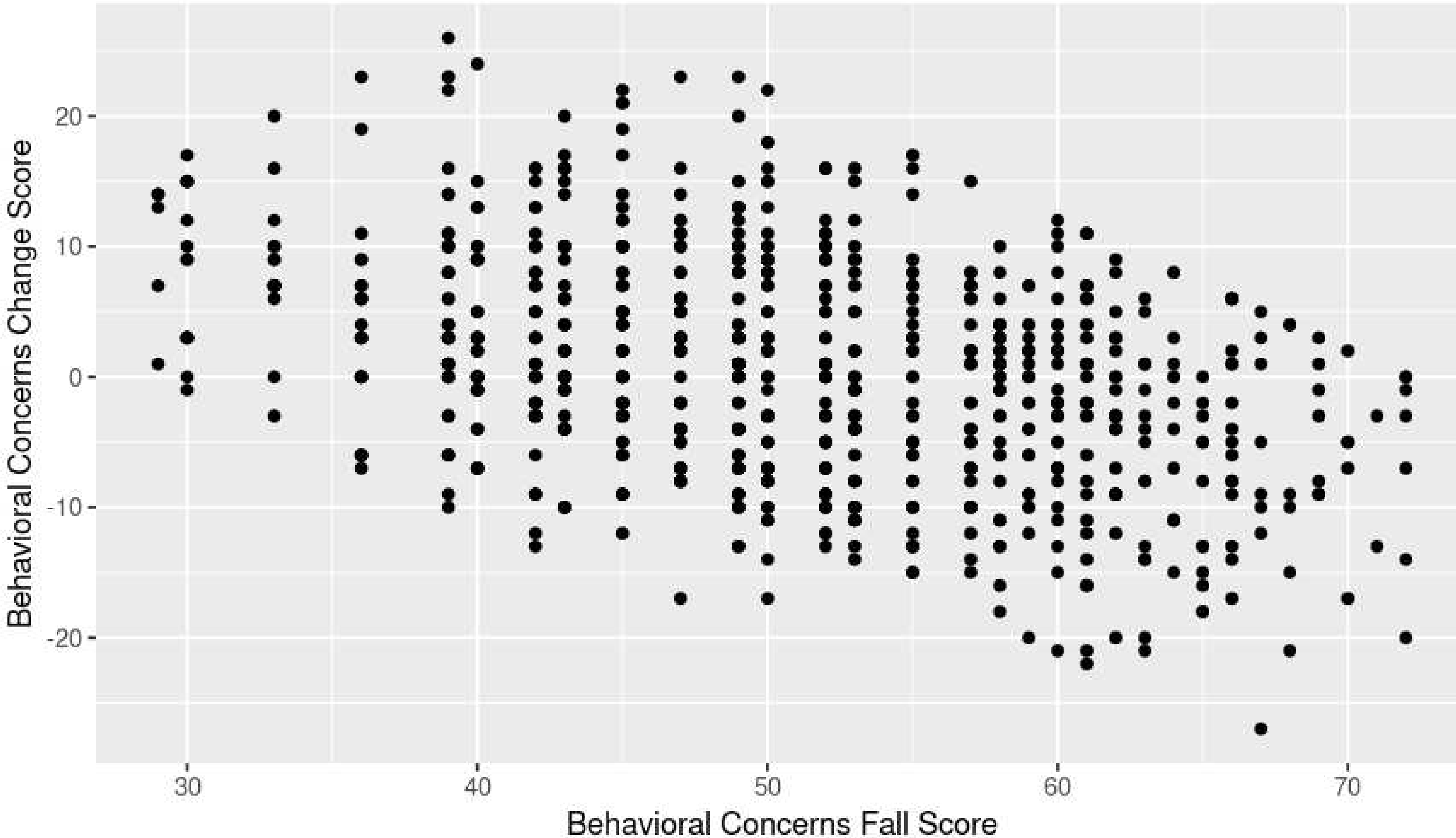
Finally, scatterplots were created using unimputed data with individual children' initial DECA scores, one each for Total Protective Factors and Behavioral Concerns. The initial outcome scores were placed on the x-axis while the change scores were placed on the y-axis. Figure 7 and Figure 8 show that children who start out with lower Total Protective Factors and Behavioral Concerns scores in Fall tend to have larger change scores from Fall to Spring.

Figure 7



Scatterplot of Child Level Total Protective Factors Fall and Change Scores

Figure 8



Scatterplot of Child Level Behavioral Concerns Fall and Change Scores

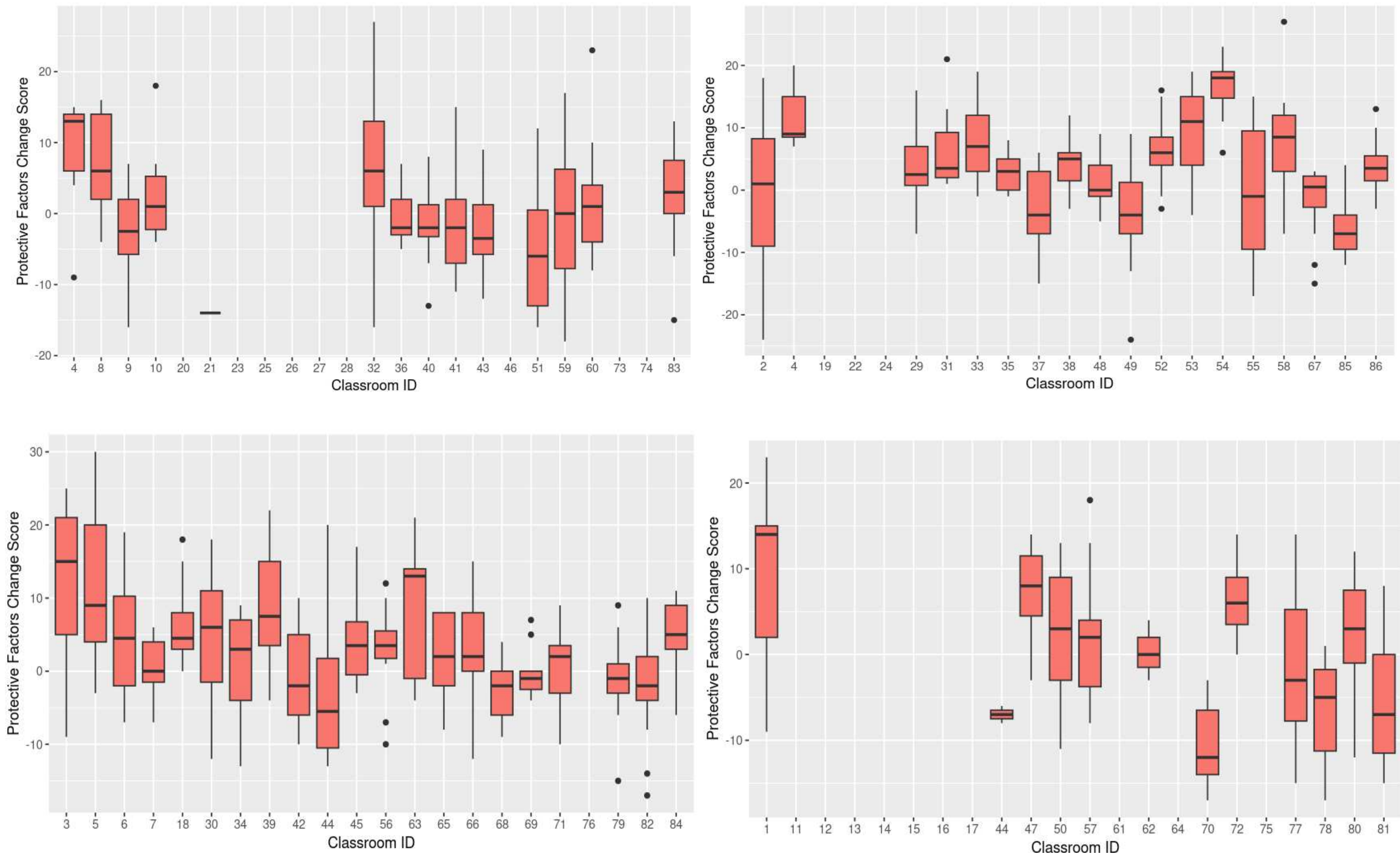
Boxplots

Figures 9 through 20 present boxplots of child level change scores for Total Protective Factors and Behavioral Concerns grouped by classroom across all quartiles of the six predictors examined in this dissertation. Across all predictors, including Emotional Support, Instructional Support, Classroom Organization, Responsive Teaching, Proactive Management and Routines, and Cognitive Facilitation, there is substantial classroom-level variability in Total Protective Factors Change Scores and Behavioral Concerns change scores. Some classrooms exhibit a wider spread of scores, indicating greater within-classroom differences in child experiences, while others display more uniform patterns. The presence of outliers across multiple classrooms suggests that some children have significantly higher or lower change scores compared to their peers in the same classroom. Although median scores do not show drastic shifts across quartiles for most predictors, the distribution of scores varies across classrooms. Certain quartiles consistently include classrooms with higher or lower Total Protective Factors or Behavioral Concerns change scores, suggesting systematic differences in instructional effectiveness, classroom management, or environmental factors. Quartiles with greater dispersion, indicated by longer whiskers and more widely spread boxplots, highlight classrooms where improvements in child outcomes from Fall to Spring are more heterogeneous.

Missing data is a pervasive issue across all predictors and outcomes, with some classrooms entirely lacking outcome data. The distribution of missing data varies by predictor. For all predictors included in this dissertation, there are different missingness patterns across quartiles. These patterns suggest that missing data may not be random and could be related to classroom characteristics, child demographics, or assessment completion rates, potentially introducing bias in model estimates. To address this issue, multiple imputation techniques and sensitivity analyses should be considered in future analyses.

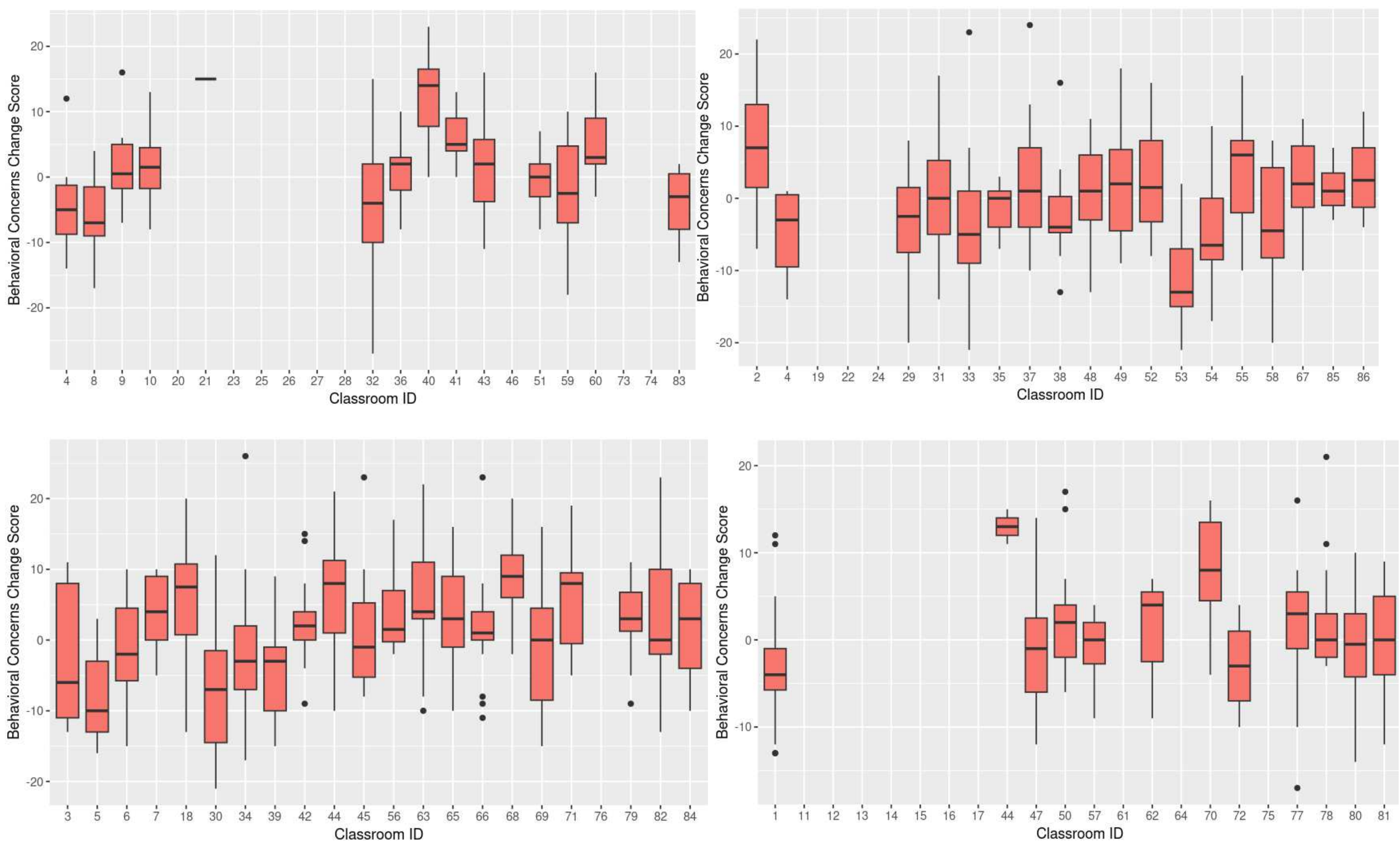
Overall, the findings indicate that classroom-level differences are evident across all predictors and outcomes, with some classrooms consistently exhibiting higher or lower Total Protective Factors and Behavioral Concerns change scores. There is notable heterogeneity in improvements in child outcomes from Fall to Spring across quartiles, suggesting that classroom quality measures such as Emotional Support and Instructional Support may have varying levels of effectiveness in different contexts. The widespread presence of missing data, particularly in specific quartiles, could impact the interpretation of results if not handled appropriately. These findings highlight the necessity of thoughtful multilevel modeling approaches, careful treatment of missing data, and robust methods for addressing variability.

Figure 9



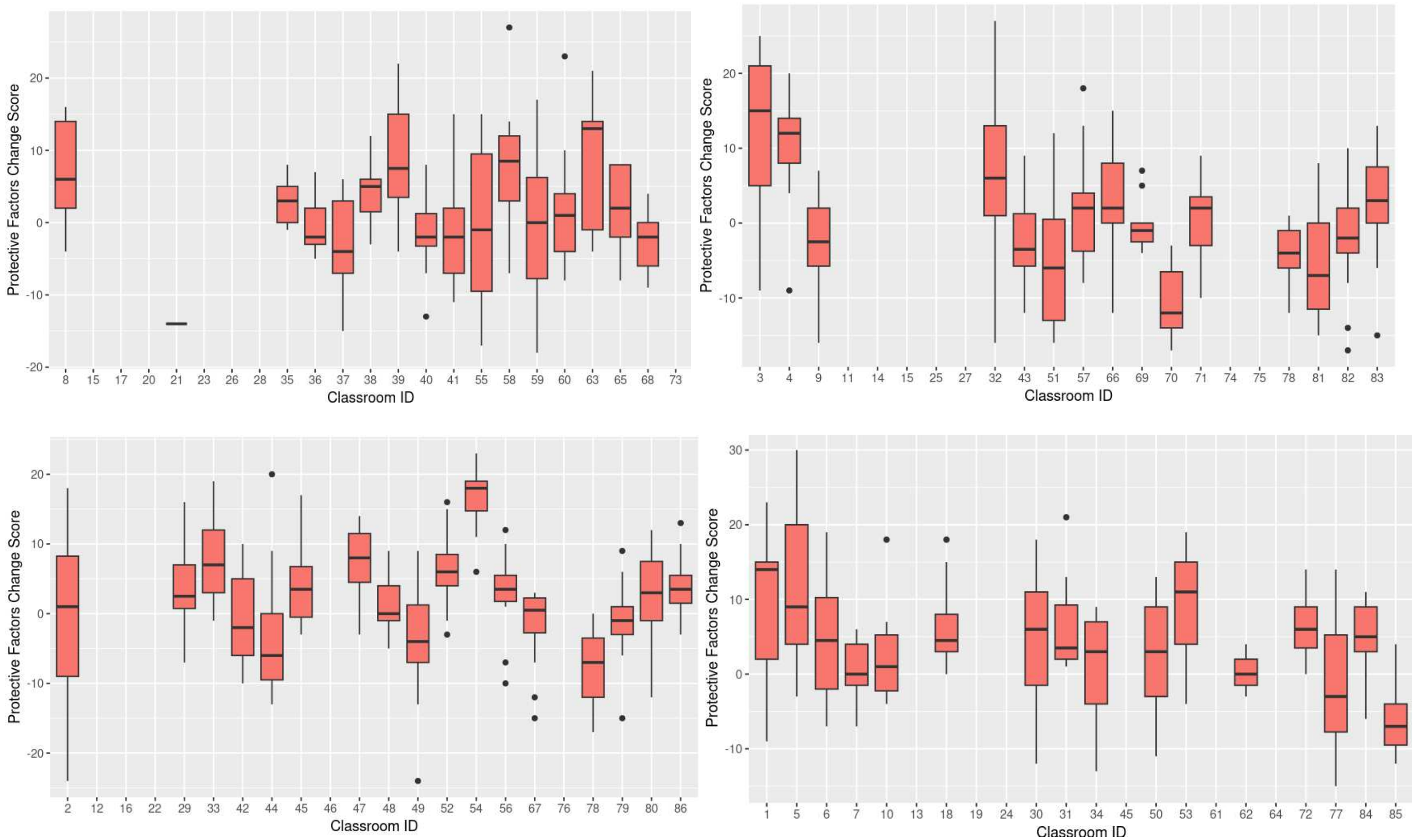
Boxplots of Total Protective Factors Change Scores Grouped by Classroom Across Quartiles of Emotional Support
Note. Quartile 1 is top left, Quartile 2 is top right, Quartile 3 is bottom left, and Quartile 4 is bottom right.

Figure 10



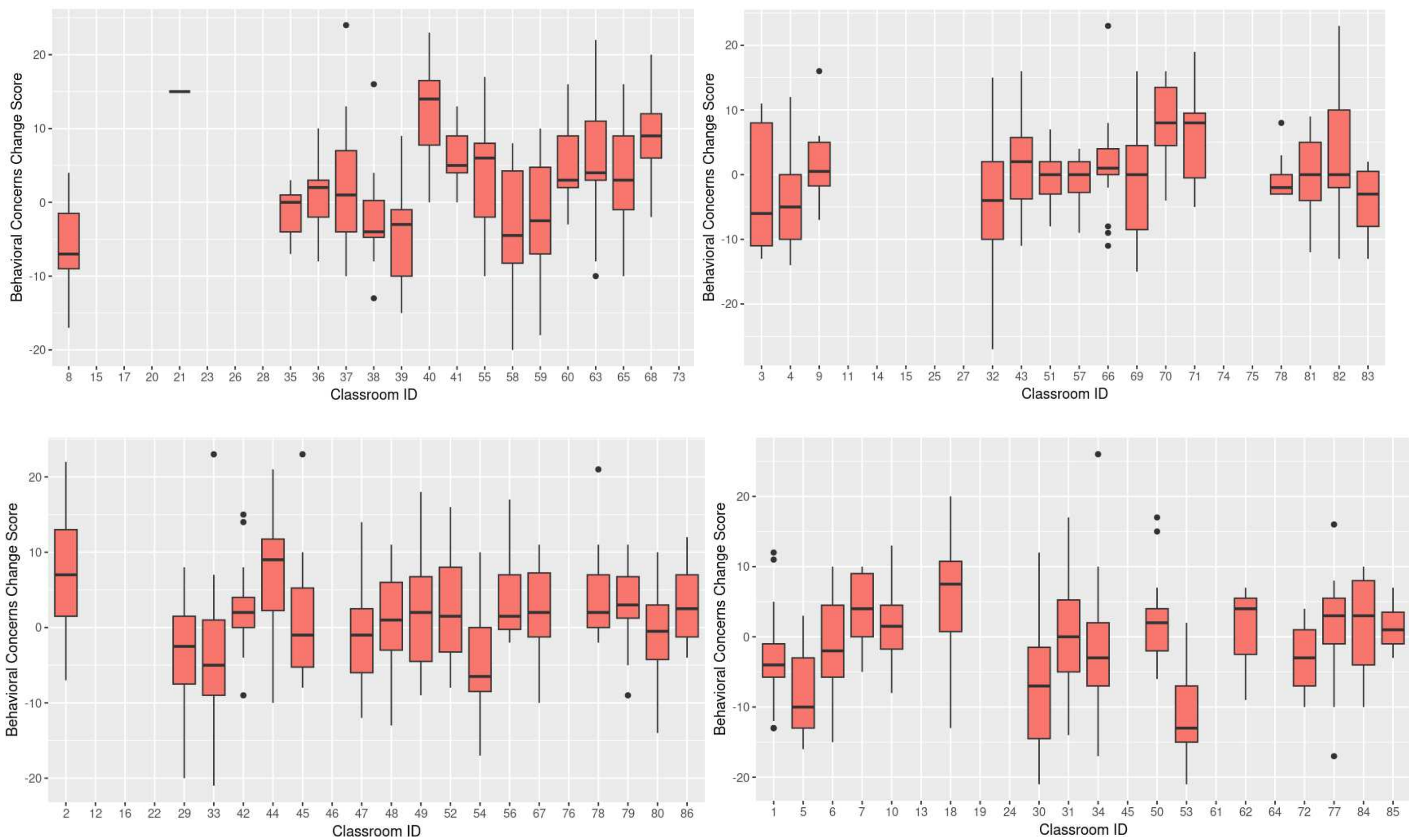
Boxplots of Behavioral Concerns Change Scores Grouped by Classroom Across Quartiles of Emotional Support
Note. Quartile 1 is top left, Quartile 2 is top right, Quartile 3 is bottom left, and Quartile 4 is bottom right.

Figure 11



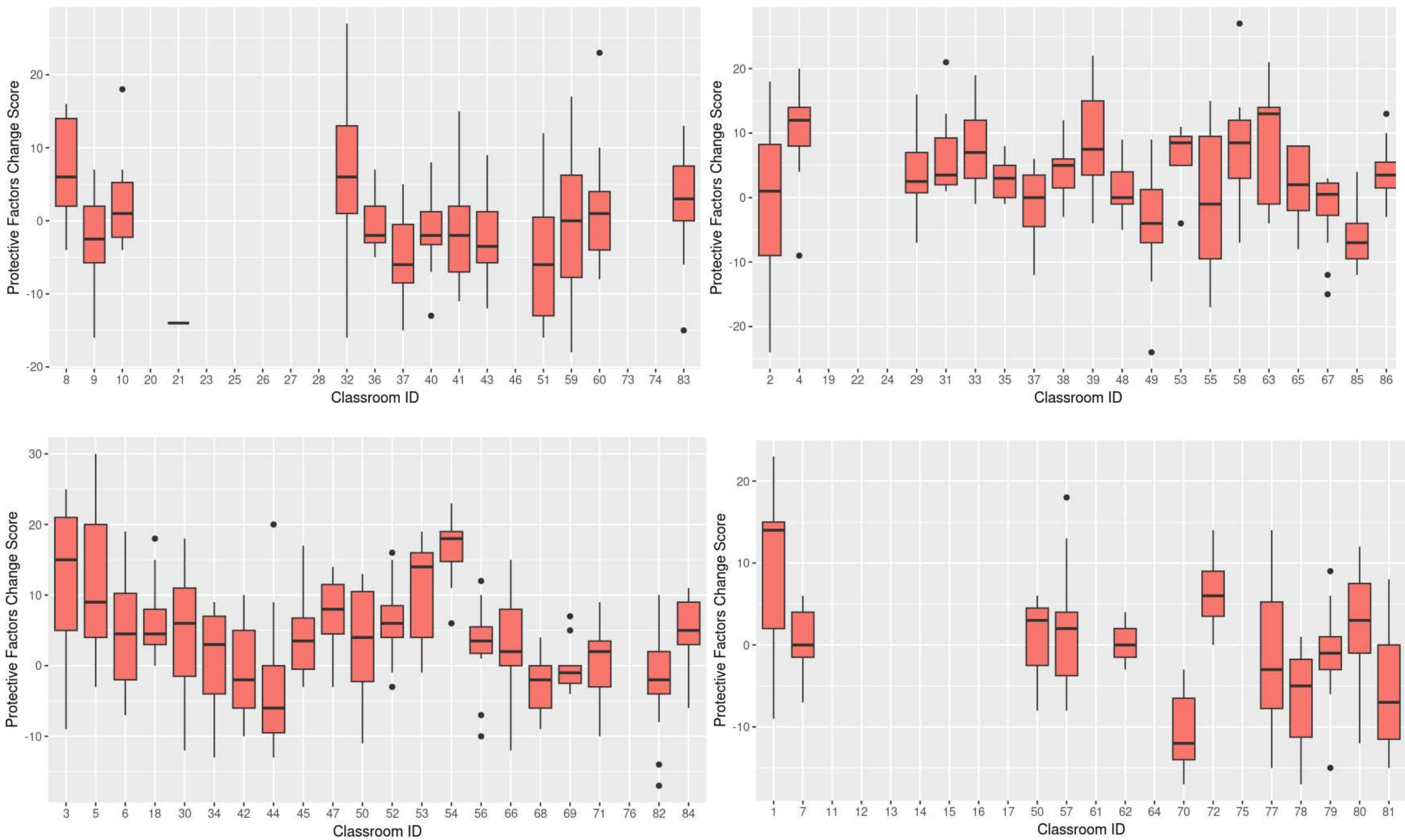
Boxplots of Total Protective Factors Change Scores Grouped by Classroom Across Quartiles of Instructional Support
Note. Quartile 1 is top left, Quartile 2 is top right, Quartile 3 is bottom left, and Quartile 4 is bottom right.

Figure 12



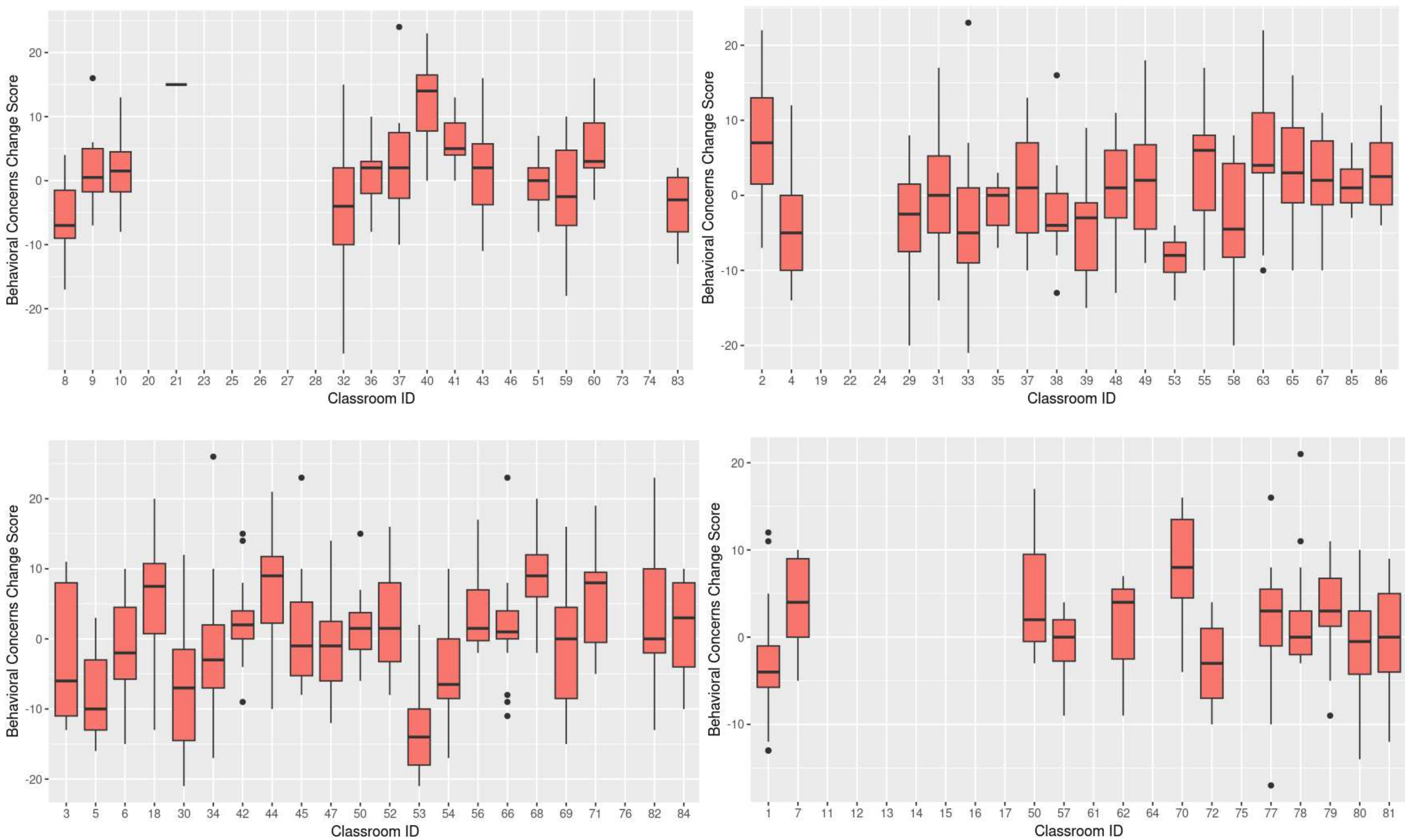
Boxplots of Behavioral Concerns Change Scores Grouped by Classroom Across Quartiles of Instructional Support
Note. Quartile 1 is top left, Quartile 2 is top right, Quartile 3 is bottom left, and Quartile 4 is bottom right.

Figure 13



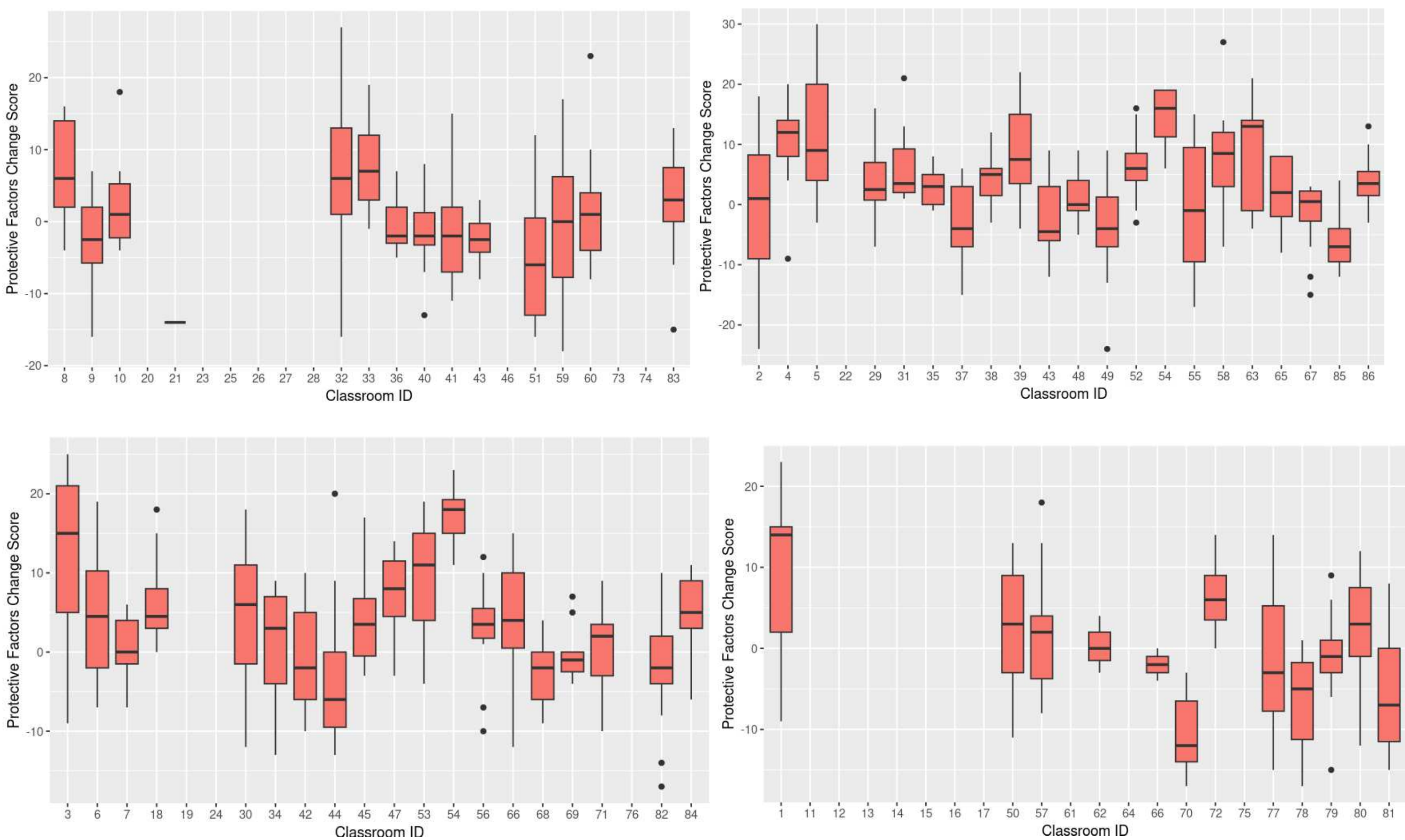
Boxplots of Total Protective Factors Change Scores Grouped by Classroom Across Quartiles of Classroom Organization
Note. *Quartile 1 is top left, Quartile 2 is top right, Quartile 3 is bottom left, and Quartile 4 is bottom right.*

Figure 14



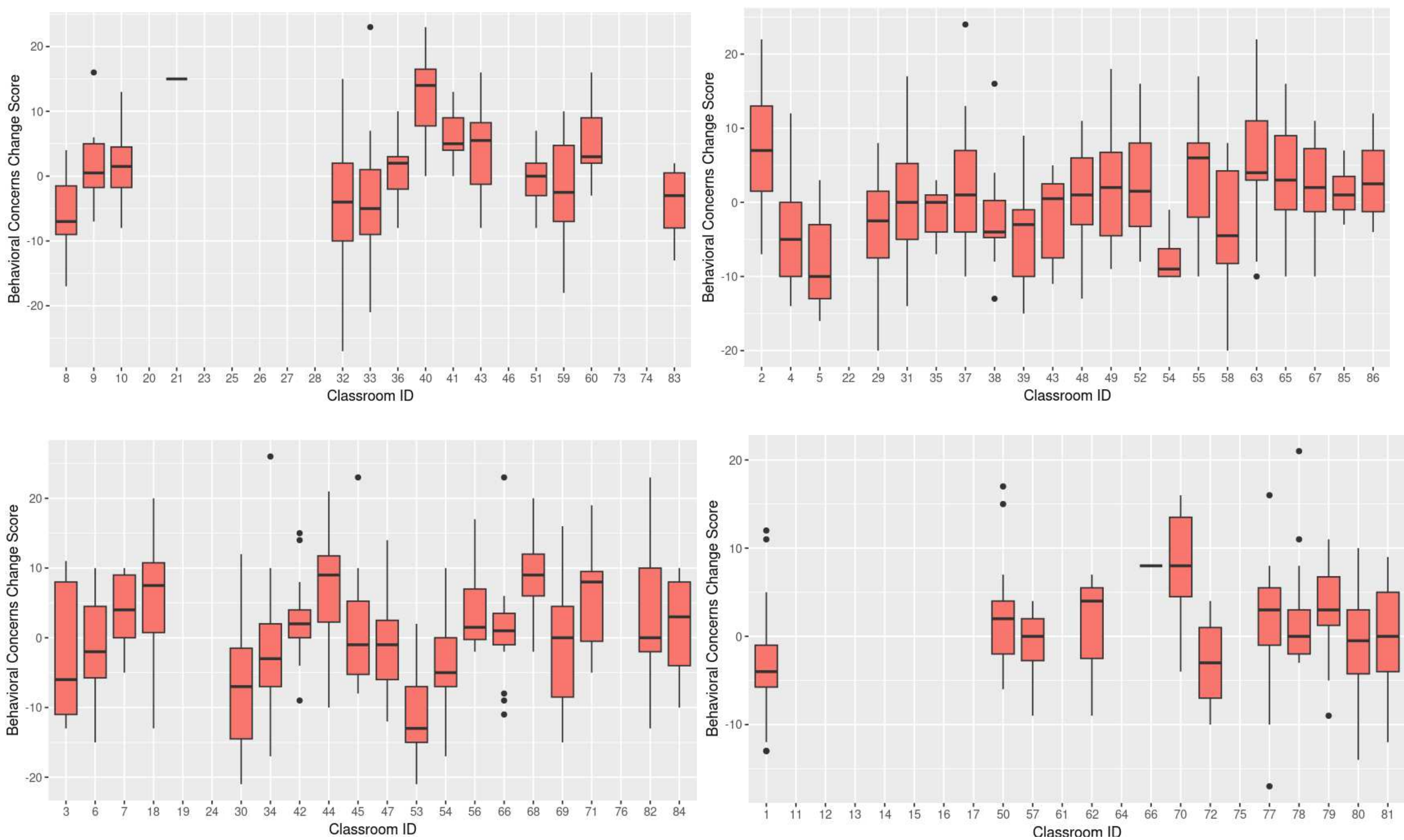
Boxplots of Behavioral Concerns Change Scores Grouped by Classroom Across Quartiles of Classroom Organization
Note. Quartile 1 is top left, Quartile 2 is top right, Quartile 3 is bottom left, and Quartile 4 is bottom right.

Figure 15



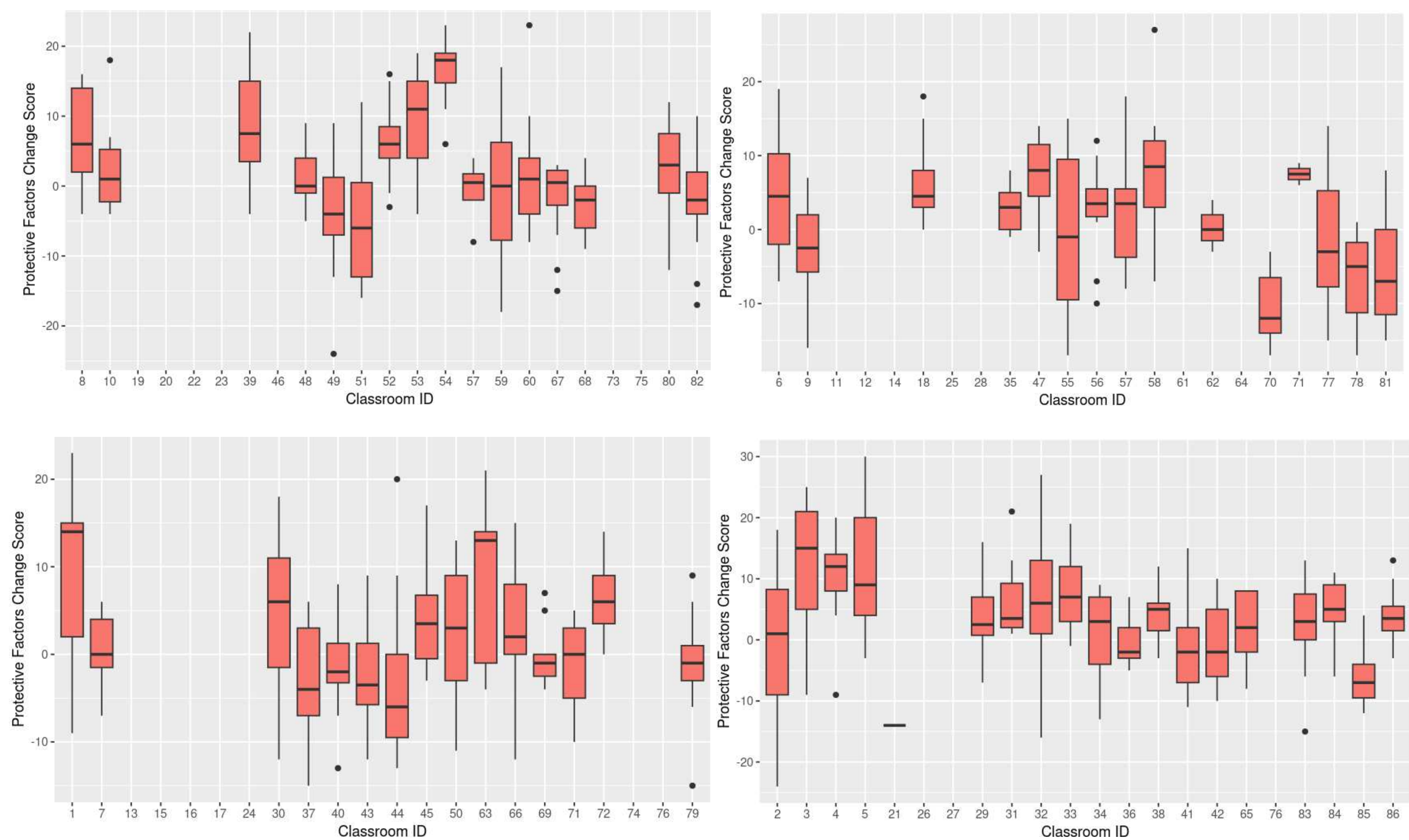
Boxplots of Total Protective Factors Change Scores Grouped by Classroom Across Quartiles of Responsive Teaching
Note. Quartile 1 is top left, Quartile 2 is top right, Quartile 3 is bottom left, and Quartile 4 is bottom right.

Figure 16



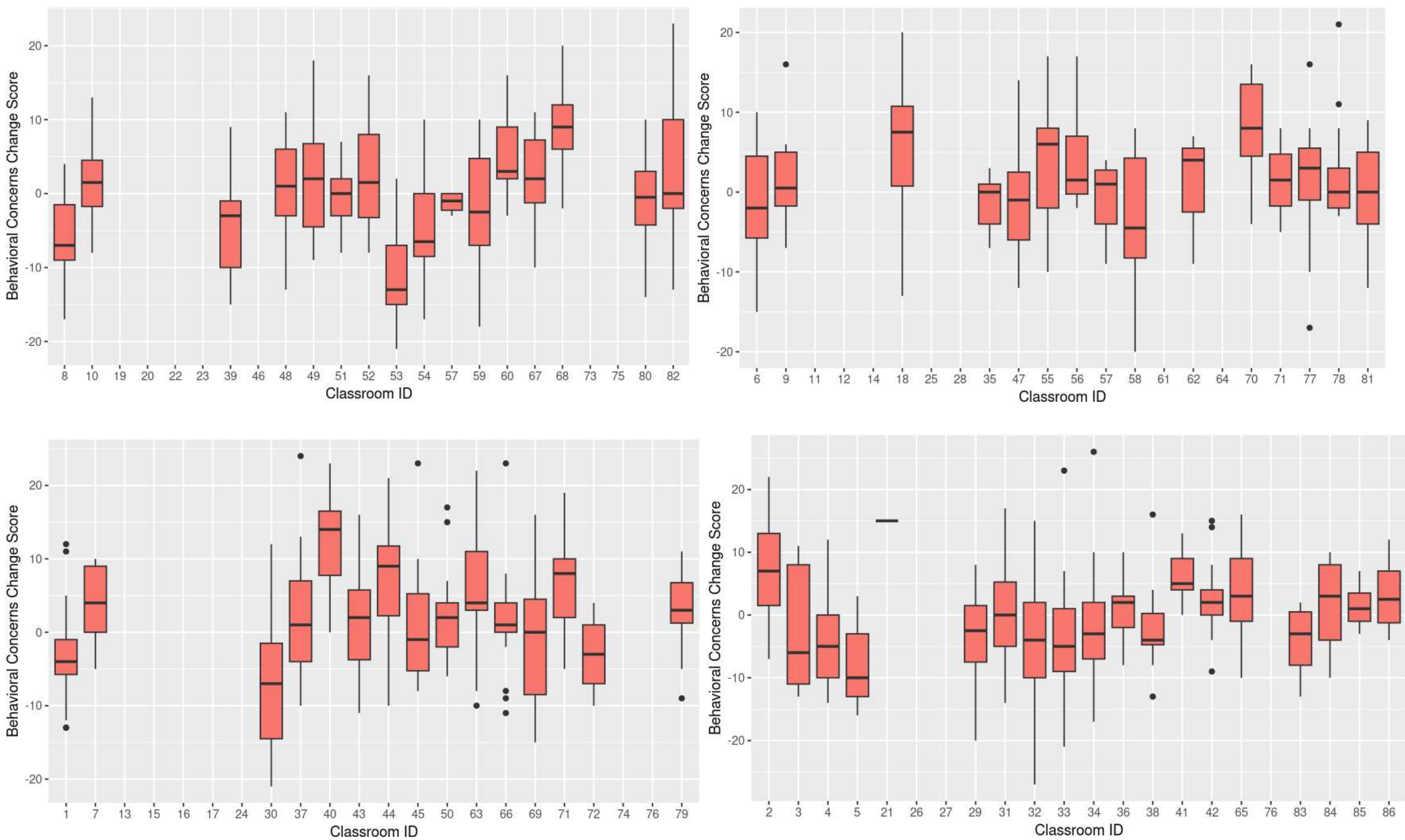
Boxplots of Behavioral Concerns Change Scores Grouped by Classroom Across Quartiles of Responsive Teaching
Note. Quartile 1 is top left, Quartile 2 is top right, Quartile 3 is bottom left, and Quartile 4 is bottom right.

Figure 17



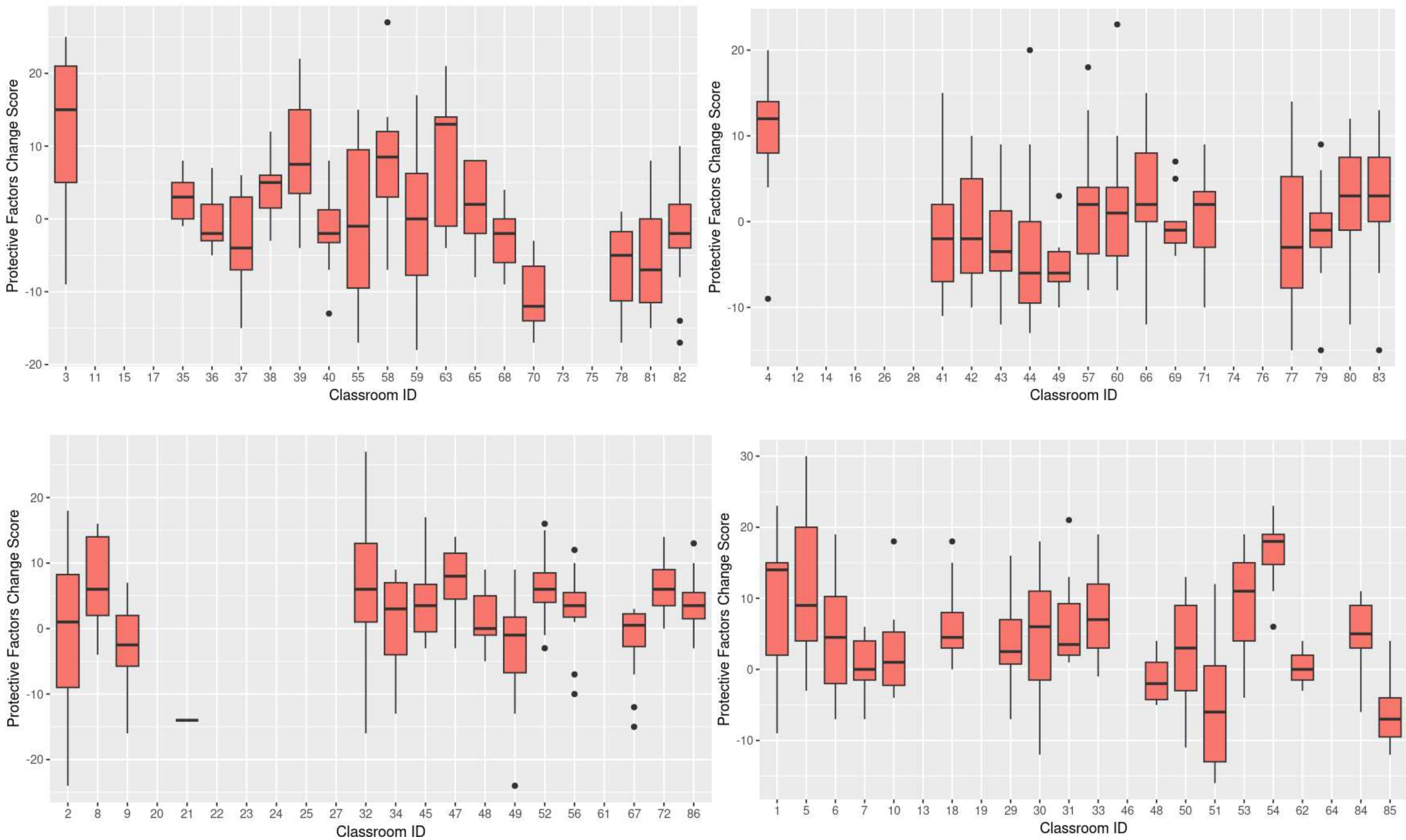
Boxplots of Total Protective Factors Change Scores Grouped by Classroom Across Quartiles of Proactive Management and Routines
Note. Quartile 1 is top left, Quartile 2 is top right, Quartile 3 is bottom left, and Quartile 4 is bottom right.

Figure 18



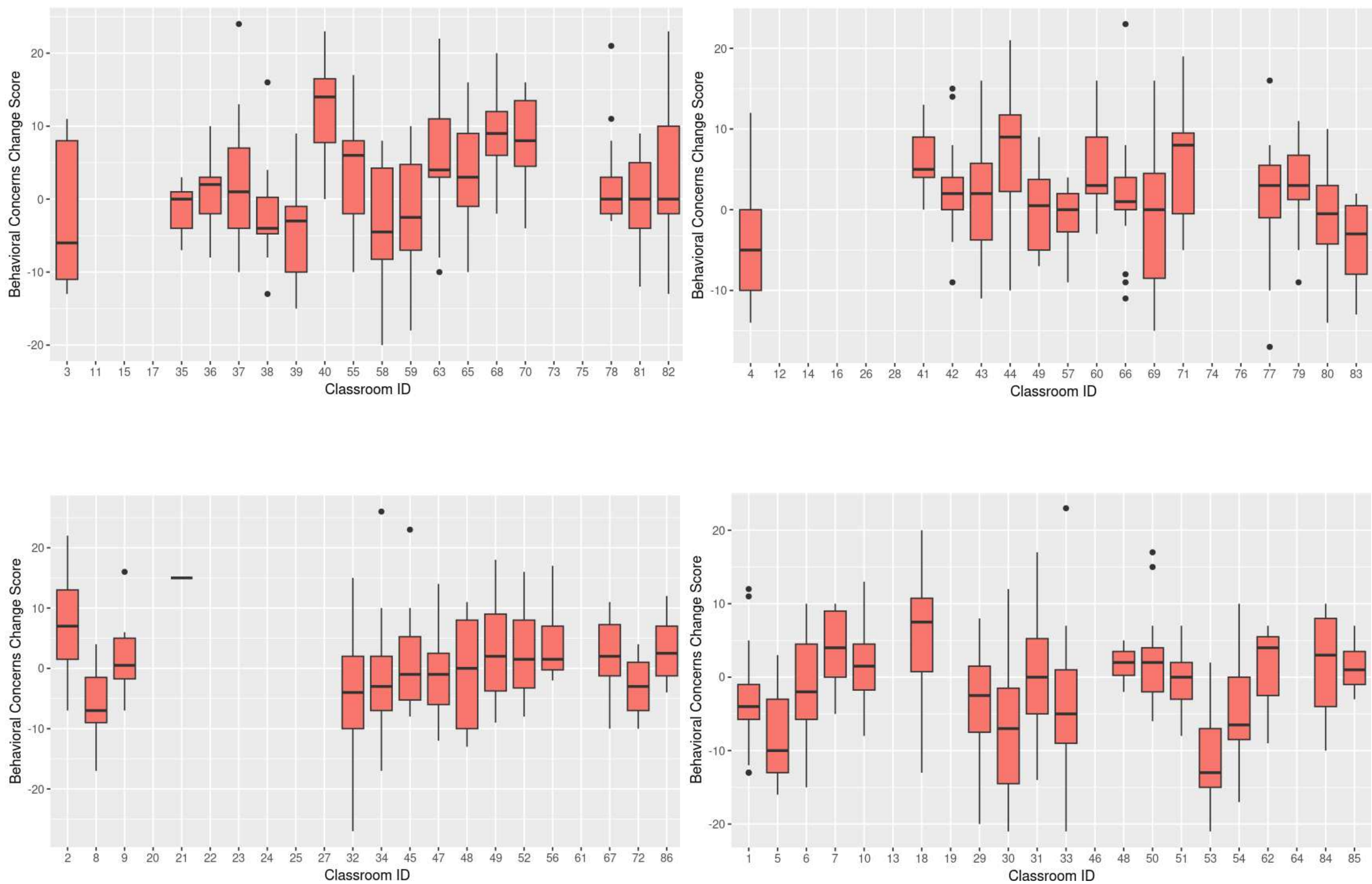
Boxplots of Behavioral Concerns Change Scores Grouped by Classroom Across Quartiles of Proactive Management and Routines
Note. Quartile 1 is top left, Quartile 2 is top right, Quartile 3 is bottom left, and Quartile 4 is bottom right.

Figure 19



Boxplots of Total Protective Factors Change Scores Grouped by Classroom Across Quartiles of Cognitive Facilitation
Note. Quartile 1 is top left, Quartile 2 is top right, Quartile 3 is bottom left, and Quartile 4 is bottom right.

Figure 20



Boxplots of Behavioral Concerns Change Scores Grouped by Classroom Across Quartiles of Cognitive Facilitation. Note. Quartile 1 is top left, Quartile 2 is top right, Quartile 3 is bottom left, and Quartile 4 is bottom right.