2019

A Multilevel Factor Analytic Investigation Of The Learning-To-Learn Scales: A More Child-Centered Look At Dimensionality

Benjamin Pratt Brumley
University of Pennsylvania, benjamin.brumley@gmail.com

A Multilevel Factor Analytic Investigation Of The Learning-To-Learn Scales: A More Child-Centered Look At Dimensionality

Abstract
Children from low-income households are at risk for entering school behind their more economically advantaged peers across major domains of school readiness. The Head Start program represents the federal government's response to these achievement gaps by mandating the use of scientifically based assessments and curricula to provide children with the necessary school readiness skills. Routine teacher-report assessment of children's school readiness using scientifically validated assessments is key to effectively guide early childhood education. Approaches to Learning is one of the five domains of school readiness targeted by Head Start. The Learning-to-Learn Scales (LTLS) is currently the only multidimensional, teacher-report assessment of Approaches to Learning that has been validated for use with Head Start students using traditional statistical methods used to identify the dimensions of the LTLS. These methods, however, do not address the multilevel nature of children nested within teacher assessors and therefore do not account for assessor variance that may compromise the validity of teacher-report child assessments. The present study applies the most advanced, multilevel factor analytic methods to examine how assessor variance impacts the validity of the LTLS dimensions. The results of this study revealed a substantial level of assessor variance was founded associated with every item of the LTLS. Accounting for assessor variance changed both the number of dimensions identified and the nature of the dimensions. Furthermore, the multilevel dimensions had greater capacity to explain variance in important external outcomes compared to dimensions identified by traditional factor analysis. The present study was the first to investigate assessor variance in teacher-report assessment of preschool-aged Head Start children. This research calls into question the validity of widely used preschool, teacher-report assessment based solely on traditional statistical methods. It, therefore, sounds an alarm to alert the early childhood education community to the need to examine assessor variance in its widely used, teacher-report assessments and where necessary use multilevel statistical methods to produce more scientifically valid assessments, especially if these assessments are used to inform decision making for young children from low-income households.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Education

First Advisor
John W. Fantuzzo

Keywords
approaches to learning, factor analysis, head start, hierarchical linear modeling, multilevel, multilevel factor analysis

A MULTILEVEL FACTOR ANALYTIC INVESTIGATION OF THE LEARNING-TO-LEARN SCALES: A MORE CHILD-CENTERED LOOK AT DIMENSIONALITY

Benjamin Pratt Brumley

A DISSERTATION

in

Education

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

_____

John W. Fantuzzo
Albert M. Greenfield Professor of Human Relations

Graduate Group Chairperson

_____

J. Matthew Hartley, Professor of Education

Dissertation Committee

Vivian Gadsden, William T. Carter Professor of Child Development

Jonathan D. Schweig, Rand Corporation

Katherine Barghaus, Executive Director, Penn Child Research Center

A MULTILEVEL FACTOR ANALYTIC INVESTIGATION OF

THE LEARNING-TO-LEARN SCALES:

A MORE CHILD-CENTERED LOOK AT DIMENSIONALITY

DEDICATION

I dedicate this work to my wife, Lauren. My wife has been there for years reading draft after draft and helping me through setback after setback. In better and worse times, my wife Lauren was dedicated to seeing this dissertation on record. I am deeply in debt to her.

ACKNOWLEDGMENT

This dissertation would not have been possible without the support of many intelligent and talented folks who helped me along the way. First, I would like to thank my wife, Lauren. Without whom, I am not sure I would have made it through this process. Being both a coach, a coauthor, a mentor, and proofreader while also being an outstanding scholar is quite an accomplishment. But that is to be expected from such an amazing person.

I would also like to thank my extended family for providing unconditional support. Trey, Jessica, Taylor, Geralyn and Scott, thank you, thank you for living through the ups and downs of this process vicariously through me. Also, thank you to the Nogays, Careys, Pevets, Campbells, Carenbauers, and Danzis who helped out in so many ways with encouragement, support and companionship.

Next, I would like to broadly thank the Penn Child Research team. This includes Staci Perlman, Heather Rouse, Whitney LeBoeuf, Katherine Barghaus, Cassandra Henderson, Kristen Coe, Erin Bogan, TC Burnett and many others. You all have been amazingly supportive critiquing and helping in a bunch of different places. Thank you all through the many years of this process. Especially thank you to Cassandra Henderson, Kristen Coe who bore the brunt of all the hard work. Thank you for helping me get to my finish line. They are two amazing colleagues and I am excited to celebrate with them when they cross their own finish line.

Next, I would thank my dedicated committee including John Fantuzzo, Katherine Barghaus, Jonathan Schweig, and Vivian Gadsden. You all have been incredibly supportive and provided your time improving and refining such an important contribution

to the research literature. Thank you for walking with me along the way. I appreciate all of your attention and hard work helping me get to this point.

Last but certainly not least, I would like to thank all of the participants and folks that usually do not get mentioned in these kinds of acknowledgements. Thank you to the unsung contributors to this process. The IT folks, the GSE administrators, the Head Start Parent Policy Council, federal grant managers, participating teachers and student in Philadelphia and many others. Without all of your support, I would not be here today. Thank you for your help, advice, and refinement. You all have helped get me to this point.

Finally, I would like to acknowledge two sources of support for this dissertation. The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B090015 to the University of Pennsylvania. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

This publication was also made possible by Grant Number 90YE0162 from the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Office of Planning, Research and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

ABSTRACT

A MULTILEVEL FACTOR ANALYTIC INVESTIGATION OF THE LEARNING-TO-LEARN SCALES: A MORE CHILD-CENTERED LOOK AT DIMENSIONALITY

Benjamin P. Brumley

John W. Fantuzzo

Children from low-income households are at risk for entering school behind their more economically advantaged peers across major domains of school readiness. The Head Start program represents the federal government's response to these achievement gaps by mandating the use of scientifically based assessments and curricula to provide children with the necessary school readiness skills. Routine teacher-report assessment of children's school readiness using scientifically validated assessments is key to effectively guide early childhood education. Approaches to Learning is one of the five domains of school readiness targeted by Head Start. The Learning-to-Learn Scales (LTLS) is currently the only multidimensional, teacher-report assessment of Approaches to Learning that has been validated for use with Head Start students using traditional statistical methods used to identify the dimensions of the LTLS. These methods, however, do not address the multilevel nature of children nested within teacher assessors and therefore do not account for assessor variance that may compromise the validity of teacher-report child assessments. The present study applies the most advanced, multilevel factor analytic methods to examine how assessor variance impacts the validity of the LTLS dimensions. The results of this study revealed a substantial level of assessor variance was founded associated with every item of the LTLS. Accounting for assessor

variance changed *both* the number of dimensions identified and the nature of the dimensions. Furthermore, the multilevel dimensions had greater capacity to explain variance in important external outcomes compared to dimensions identified by traditional factor analysis. The present study was the first to investigate assessor variance in teacher-report assessment of preschool-aged Head Start children. This research calls into question the validity of widely used preschool, teacher-report assessment based solely on traditional statistical methods. It, therefore, sounds an alarm to alert the early childhood education community to the need to examine assessor variance in its widely used, teacher-report assessments and where necessary use multilevel statistical methods to produce more scientifically valid assessments, especially if these assessments are used to inform decision making for young children from low-income households.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

A National Crisis in Education

In 1966, the Department of Education released the 'Coleman Report.' This report was commissioned by Congress to evaluate the equality of educational opportunities for children in the United States. Findings indicated that children from families of low income and ethnic minority status were at-risk to remain academically behind their middle class, white peers throughout their school years. Over four decades of studies have documented that these achievement gaps have not been reduced (Vanneman, Hamilton, Baldwin, & Rahman, 2009). In fact, gaps between students from the highest and lowest household incomes have grown in magnitude. These gaps are now twice as large as those gaps between children from families of majority and minority racial backgrounds (Reardon, 2011). These troubling findings indicate that intervention is still needed for children from economically disadvantaged backgrounds. Moreover, recent studies indicate that children from low income families start school behind their peers and are subsequently at risk to remain behind throughout their education (Duncan et al., 2007; Karoly, Kilburn, & Canon, 2005).

Similar to the pattern identified in the Coleman Report, gaps between children are expanding over their time in school. Duncan and Magnuson (2011) documented that kindergarteners from low-income households had mathematics and reading scores that were over a standard deviation lower than children from high income households. The size of these gaps continued to increase by nearly 15% at the end of primary school (Duncan & Magnuson, 2011). This persistent gap in academic skills has led many

scholars to call for a more comprehensive and intensive national response to intervene for children living in poverty before they enter kindergarten (Barnett, 2011).

The national conversation around achievement gaps now reflects a body of empirical literature that shows gaps exist beyond tests of academic achievement to other domains of children's school readiness functioning (Shonkoff & Phillips, 2000; Duncan et al., 2007). When looking across multiple domains of readiness for school, studies have documented that children from low-income households are also behind their peers from high-income households on important non-academic skills (Issacs & Brookings, 2011). Today, the discrepancy across these multiple skills is referred to as a 'gap in school readiness' (Duncan et al., 2007).

School readiness is a construct that represents five domains of early learning skills that have been shown to be important for academic success in the classroom (Administration for Children and Families [ACF], 2015a; National Education Goals Panel [NEGP], 1995). These domains include cognitive abilities, language, literacy knowledge, physical development, social-emotional competencies, and approaches to learning. Children from low income households not only enter behind their peers from higher-income households in mathematics, reading, and motor skills they also enter school less proficient on social-emotional skills and approaches to learning abilities (U.S. Department of Education, 2002; Zill & West, 2001). The gaps across critical developmental skills stress that these young children need holistic intervention rather than solely reading and mathematics support before they enter kindergarten.

Head Start a National Response

Since 1965, the national Head Start program has been the primary federal early childhood intervention for young children living in poverty. From the very beginning, Head Start has been shaped by a comprehensive developmental-ecological theoretical framework and rigorous empirical research to inform the implementation of its theory of change and to evaluate its effectiveness (Zigler & Styfco, 2010). With this guide, Head Start has grown significantly in size and scope since its introduction as part of President Johnson's War on Poverty (Zigler & Styfco, 2010). It was originally designed as an eight-week summer intervention program, but today it is the largest federally funded early childhood program serving 1,060,620 children and their families across 1,654 programs in every state in America (ACF, 2015b). Head Start, with an operating budget of nearly 8.5 billion dollars, continues to be the federal government's primary intervention for preschool children living in low income households (ACF, 2015b). With the Reauthorization of Head Start in 1998, the mission of Head Start shifted from a focus on enhancing children's general social competencies to a more explicit goal of promoting the *school readiness* of young children (Zigler & Styfco, 2010).

A Developmental Ecological Model Guides Head Start

The developmental-ecological model has shaped Head Start's approach to supporting young children and their families since its inception (Bronfenbrenner & Morris, 2006). This theoretical framework posits that children's developmental growth is shaped by interactions between the child's *Person* characteristics and their environmental *Context* over *Time*. The first component of the model through which Head Start views the developing child is their unique *Person* characteristics that play a role in their burgeoning

readiness for school. Head Start seeks to support the "whole child" by targeting a holistic set of five domains of children's functioning that are important for school readiness: cognitive abilities, language, literacy knowledge, physical development, social-emotional competencies, and approaches to learning. Head Start underscored its commitment to supporting children across these five major domains of school readiness by establishing performance standards to guide intervention (i.e., *Head Start Learning Outcomes Framework*; ACF, 2015a; National Education Goals Panel, 1995). Thus, Head Start recognizes the need to support children's development of school readiness in the broadest sense.

The second important component of the developmental-ecological model is the *Context* in which the developing *Person* learns and interacts with others (Tudge et al., 2016). The most proximal and influential contexts for young children are their home and classroom environments (Tudge et al., 2016). As such, Head Start is a two-generational program that seeks to promote school readiness competencies at home and in the classroom. Head Start most directly influences the classroom context to intentionally influence children's development of school readiness. Head Start's programmatic efforts are intentionally "child-centered" in which dynamic classroom environments are designed to promote child-centered learning and to individualize their approach to enhance the school readiness competences of whole child. This is accomplished through assessing each child's profile of competencies and implementing developmentally-appropriate curricula and teaching practices that are designed to help develop these competencies for school entry (Bierman, Domitrovich, Nix, Gest, Welsh, Greenberg ... & Gill, 2008).

Finally, *Time* plays an important role in the developmental-ecological model and informs Head Start's mandates. *Time* includes both the extent to which activities occur consistently over time in the child's learning environment as well as monitoring children's progress across time (Bronfenbrenner & Morris, 2006). The latter concept is evident in Head Start's use of routine, ongoing assessment to monitor children's development of school readiness skills over time (ACF, 2014). In fact, Head Start mandates that teachers assess children's functioning across all school readiness domains at least three times over the course of the school year (ACF, 2014). Head Start's emphasis on routine assessment recognizes that children's school readiness skills are constantly changing and therefore should be monitored at regular intervals to advance these important competencies. The results of such assessment are used to tailor the teaching and learning environment to meet the individual needs of each child. Individualized instruction across time represents an integration of a focus on *Person*, *Context*, and *Time*, and is the pinnacle of Head Start's approach to effectively prepare children from low-income households to be ready for school (ACF, 2014).

The Role of Research to Evaluate Head Start

For the past five decades, researchers have been investigating the effectiveness of Head Start. The first major evaluation, known as the 'Westinghouse Study' concluded that Head Start boosted children's intelligence in the short term, but the evaluation did little to demonstrate program effectiveness for school readiness competencies (Cicirelli, 1969; Ramey & Ramey, 2004). Over the following decades, 38 additional studies indicated that participation in Head Start showed weak to moderate positive effects on school readiness outcomes including language, cognitive development, social-emotional

competencies (What Works Clearinghouse [WWC], 2015; Zigler & Styfco, 2010). These studies indicated the promise of the program, but focused only on short-term outcomes, which did little to reassure policy makers of the lasting benefits of Head Start.

At the request of congress in the late 1990s, the U.S. General Accounting Office (GAO) reviewed these early evaluations and concluded that they provided inadequate evidence from which to draw conclusions about the impact of Head Start. They were not of sufficient methodical quality from which to base decisions about the future of Head Start. The GAO reported that the majority of these studies were "too old, methodologically weak, or statistically problematic" to support a conclusion about the effectiveness of Head Start (Ziger and Styfco, 2010). A more rigorous methodological design was needed to provide more disciplined evidence about the effectiveness of Head Start. In 2000, Congress allocated funding for the first randomized controlled trial of Head Start, the "Impact Study." This study was designed to meet rigorous, contemporary methodological standards (WWC, 2015).

The Impact Study involved a nationally representative sample of approximately 5,000 preschool children who enrolled in the program at the age of three or four years old. Overall, the findings from the Impact Study indicated that Head Start demonstrated weak to moderate effects on school readiness outcomes in preschool (Puma et al., 2012; Barnett et al., 2011). The effects were examined separately for three-year-olds and four-year-olds to see if there were stronger effects for different age groups of children entering Head Start (Puma et al., 2012). Findings demonstrated that children benefitted more from the program if they enrolled at the age of three years old compared to four years old. In particular, three-year-olds demonstrated much greater positive effects across all major

domains of school readiness compared to four-year-olds (Puma et al., 2012). However, by kindergarten, initial gains among four-year-olds had disappeared and began to fade for three-year-olds. The three-year-olds demonstrated small continued benefits for social-emotional skills in kindergarten, and, by first grade, they performed marginally better than the control group in only one school readiness skill—Oral Comprehension (Puma et al. 2012). When followed-up in third grade, all effects of Head Start disappeared for both age groups.

Scholars largely concluded that the Impact Study demonstrated relatively weak effects of Head Start signaling the need for significant program reform (Barnett, 2011; Mead, 2014). The research community recommended that Head Start needed to focus more explicitly on its mechanisms for achieving school readiness outcomes (Barnett, 2011). They advocated for Head Start to develop a more evidenced-based logic model for effective intervention of the most strategic school readiness competencies (Mead, 2014; ACF, 2012, p. 8). This would require moving Head Start into a more intentional, data-based decision-making culture and a more intentional focus on "the few and the powerful." That is, major domains of child functioning that are most predictive of school readiness and are most likely to produce robust early learning trajectories that will significantly narrow achievement gaps. The advisory committee appointed by the Secretary of the U.S. Department of Health and Human Services recommended "implement[ing] the strongest and most current evidence-based practices" to increase the longevity of Head Start's effectiveness for young children (Head Start Research and Evaluation Advisory Committee, 2012).

The Reauthorization of the Head Start Act in 2007

The evaluations of Head Start reinforced Congress's resolve to improved Head Start by mandating an explicit directive to use scientifically based evidence to improve Head Start's effectiveness (P.L. 110-134, 2007). The introduction of science to the language of Head Start's reauthorization meant that Head Start providers must now use practices based on scientifically disciplined evidence. According to the reauthorization, research that applies "rigorous, systematic, and objective methodology to obtain reliable and valid knowledge" produces the scientific evidence to inform Head Start practices (P.L. 110-134, 2007). As such, Congress called on Head Start to use scientifically based assessment to (a) support instruction, (b) evaluate the extent to which programs are addressing the needs of the community, and (c) inform professional development plans. As part of the reauthorization, Head Start must implement assessments that have undergone intensive scientific review for the student populations served. These new mandates raised the bar of methodological rigor for assessments that Head Start must use.

A previously proposed assessment system for Head Start, known as the National Reporting System (NRS), failed to meet the new directive to use science and quality evidence to improve Head Start's effectiveness. The NRS was comprised of measures that lacked scientific support to validate their use in Head Start (Meisels & Atkins-Burnett, 2004). They were "rife with class prejudice and not developmentally appropriate" (Meisels & Atkins-Burnett, 2004). As such, these assessments were widely criticized for their psychometric proprieties and lack of validity evidence (National Research Council [NRC], 2008, p. 53). They had not been examined for statistical bias or piloted to establish their validity. With the new mandates of scientifically based

assessment, they were not acceptable to use in assessing Head Start children across the nation. Over 200 researchers, educators and practitioners signed letters to Congress indicating their concerns about the need for valid measurement capable of effectively guiding the educational practice with the diverse Head Start population (NRC, 2008). In response, Congress commissioned the National Research Council to identify appropriate scientific assessment available for use with young children (NRC, 2008). The specific charge of the NRC was to "was the identification of important outcomes for children from birth to age [five] and the (psychometric) quality and purposes of different techniques and instruments for developmental assessments" (p. 2 NRC, 2008).

The National Research Council Report

In response to this charge, the NRC committee created guidelines to judge the scientific integrity of early childhood assessments. The NRC also identified 'widely available' measures by school readiness domains and included them in the report appendices. However, they did not apply the scientific guidelines to the list of widely available measures they identified. This was a *major shortcoming* of this report (Barghaus & Fantuzzo, 2014).

The quality guidelines put forth by the NRC were largely drawn from the existing *Standards for Educational and Psychological Testing* sources of validity evidence (APA, AERA, & NCME, 2014). The Standards are used ubiquitously in education and psychological testing because they reflect current psychometric theory and research on what constitutes valid and reliable assessment practices (Camilli, 2006). The NRC reported on five sources of validity evidence, as indicated in the Standards, that are necessary for valid assessment (NRC, 2008, pp. 192-195). The NRC documented that

there must be clearly documented evidence of (a) instrument content, (b) the response process, (c) the internal structure, (d) relations to other variables, and (e) the consequential validity for all early childhood measures to be deemed scientifically based assessments. All five sources of validity evidence are needed to ensure valid assessment.

The first three types of validity evidence are based on the instrument content, response process and internal structure. All three are needed to ensure an internally valid instrument. First, 'Evidence Based on Instrument Content' is derived from the systematic process used to develop and evaluate the targeted construct's definition and corresponding items (NRC, 2008, p. 192; Downing & Haladyna, 1997; Kane, 2006). It ensures that all content from the targeted domain has been representatively sampled from the research knowledge-base (Downing, 2006). Second, 'Evidence Based on the Response Process' comes from documentation of the extent to which "all sources of error associated with test administration are controlled" (NRC, 2008, p. 192; Downing, 2003). For example, in teacher-report assessments, error associated with test administration can include teachers' knowledge of the target construct(s) assessed by the items, ambiguous wording of items that is interpreted differently by respondents, and teachers' skills for observing children and accurately applying the evaluation criteria (Downing, 2003). Evidence that these sources of error have been mitigated comes from documentation of systematic, scientific development of the instrument and an evaluation of test users' training on how to accurately use the instrument (Downing, 2003). Third, 'Evidence Based on Internal Structure' of an assessment refers to the extent to which there is evidence that items measure the targeted constructs for its intended use (NRC, 2008, p. 193). Assessments that aim to capture multiple constructs should have evidence from the

most scientifically advanced factor analytic methods supporting their dimensionality (Gorsuch, 2003).

The two remaining types of validity evidence include that which is based on the relations to other variables (i.e., external validity) and evidence of consequential validity. In addition to the first three sources of validity evidence, a quality measure must also demonstrate 'Evidence Based on Relations to Other Variables', otherwise referred to as external validity (NRC, 2008, pp. 193-194). This evidence is provided through documentation that the assessment's constructs are appropriately correlated with other independent measures that have been validated for use with the target population. Finally, 'Evidence Based on the Consequences of Using an Assessment Instrument' must be provided. Consequential validity for child assessments is demonstrated when a measure can be practically used by professionals to assess children's growth and development. Therefore, these scientifically based guidelines for the assessment of young children established a comprehensive basis for the evaluation of currently available assessments that requires producing multiple types of validity evidence to warrant use.

Next, the NRC (2008) reviewed 'widely-available' assessments. These widely-used measures were organized by the five school readiness domains from the National Education Goals Panel (1995). The report indicated that all five readiness domains had several widely-used assessments that were available to early childhood providers. There were 22 assessments of Language and Literacy available to assess children's progression in language competencies, 14 assessments were available to document children's cognitive abilities, 21 assessments for children's social-emotional competencies, 17 instruments to assess children's Physical Well-Being and Motor Development, and 11

assessments for components of Approaches to Learning – 5 of which were used by teachers to report on children's classroom behaviors (i.e., "teacher-report instruments"). The report suggested that these measures provided the assessment capacity to monitor at least sub-components of children's progress on these school readiness domains.

In the report, the NRC reiterated that there was an overwhelming body of evidence that Approaches to Learning was a complex multi-faceted construct that was essential for early school success (NRC, 2008). However, in contrast to the other four domains, there was only one, multidimensional assessment identified that captured distinct multiple dimensions of this domain with validity. This assessment was the *Preschool Learning Behaviors Scale* ([PLBS]; McDermott, Leigh, & Perry, 2002).

The Importance of Approaches to Learning

Approaches to Learning (ATL) is a multidimensional domain of school readiness that represents skills that connect young children behaviorally, emotionally, and cognitively to the learning process (Fantuzzo, Perry & McDermott, 2004; U.S. Department of Health and Human Services, 2010; Administration for Children and Families [ACF], 2015a). This domain recognizes that preschool children are active agents in their own learning and development (Bronfenbrenner & Morris, 2006; Hyson, 2008). As such, ATL includes children's curiosity, initiative and creativity in the classroom as well as their ability to self-regulate emotion, behavior, and cognitive processes (Blair & Diamond, 2008; McClelland, Acock, & Morrison, 2006; McDermott et al., 2011). Federal agencies and all state-funded preschool programs studied by the National Institute for Early Education Research [NIEER] now explicitly mandate early childhood standards including ATL as an essential component of their school readiness goals, which

aim to empower children with the tools they need to succeed in the classroom (NIEER, 2017; U.S. Department of Health and Human Services, 2015).

Head Start created four distinct categories of ATL skills (ACF, 2015a). The *Learning Outcomes Framework* labels these as four "categories" of ATL as: (a) Emotional and Behavioral Self-Regulation and (b) Cognitive Self-Regulation (Executive Function), (c) Creativity, and (d) Initiative and Curiosity. These categories are based on research documenting relations between these domains and later learning and development (ACF, 2015a). A growing body of evidence links skills of each of the ATL categories with positive outcomes in early schooling and success in adolescence and adulthood.

Children's competencies in Emotional and Behavioral Self-Regulation reflect children's abilities to control their behavior and emotions in voluntary and adaptive ways (Calkins & Fox, 2002; Eisenberg & Spinrad, 2004). The ability to self-regulate enables children to persevere through difficult situations and deal with upsetting events (Howse, Calkings, Anatopolous, Keane & Shelton, 2003). Multivariate models connect teacher ratings of emotional and behavioral self-regulation to primary school academic achievement controlling for child and family-level characteristics (Howse, Calkings, Anatopolous, Keane & Shelton, 2003; Graizano, Reavis, Keane, & Calkins, 2007; Trentacosta & Izard, 2007). Longitudinal studies demonstrate that self-regulation skills in early childhood are predictive of long-term outcomes (Zelazo & Carlson, 2012). Children with higher behavioral regulation in early childhood tend to have higher ratings interpersonal competence and frustration tolerance in adolescence (Mischel, Shoda, & Rodriquez, 1989; Shoda, Mischel, & Peake, 1990). Children who are living in low

income households tend to be exposed to higher levels of chronic stress, which can disrupt developing self-regulatory skills (Evans & Kim, 2013). Fostering self-regulatory abilities within the Head Start setting is critical to improve outcomes for vulnerable children.

Cognitive Self-Regulation represents children's abilities to control their attention and interact with retained information (Zelazo & Carlson, 2012). These skills enable children to focus their attention, control impulses, and demonstrate flexibility in thinking and behavior (Hyson, 2008; ACF, 2015a). Empirical evidence links cognitive self-regulation and academic achievement, particularly in mathematics (Duncan et al., 2007). Preschool cognitive self-regulation also predicts verbal comprehension and mathematics above and beyond cognitive ability (i.e., McClelland, Morrison & Holmes, 2000; McClelland et al., 2007). Children who demonstrate better cognitive self-regulation in preschool have higher SAT scores, are less likely to use recreational drugs, and have a decreased likelihood of a criminal conviction as an adult (Ayduk et al., 2000; Moffitt et al., 2011; Zelazo & Carlson, 2012). Cognitive self-regulation is particularly relevant for children from low-income households. Among children from low income families, cognitive self-regulation accounts for up to 40% of the variance in standardized test scores (Waber, Gerber, Turcios, Wagner, & Forbes, 2006).

In addition to emotional, behavioral, and cognitive self-regulation, the Head Start Learning Outcomes Framework also highlights Creativity as an essential category of ATL. Creativity refers to developmentally appropriate indicators of thinking, communicating, and playing in creative and flexible ways. This includes children's capacity to ask novel questions in learning activities, demonstrate creative problem

solving, and use their imagination when playing (ACF, 2015a). Particularly in preschool, creativity is significantly related to performance on tests of both mathematics and language abilities (Holmes, Romeno, Ciraola, & Grushko, 2014). Creativity can involve exploratory and imaginative play, as well as abilities to engage in divergent thinking. These creative processes are associated with academic achievement in early childhood and are linked to long-term outcomes (Hendrick, 2001; Kaufman, Plucker, & Baer, 2008). Interventions targeting preschool children's creativity through play-based interventions show improvements in academic test scores in primary school, as well as lower rates of juvenile delinquency in adolescence (Schweinhart, & Weikart, 1998; Marcon, 2002; Hammond, Skidmore, Wilcox-Herzog, & Kaufman, 2013).

Initiative and Curiosity is the fourth and final category of Head Start's ATL framework. Children with well-developed initiative and curiosity skills demonstrate abilities to work independently, seek out new information, and demonstrate an eagerness to learn (ACF, 2015a). There has been relatively less research examining the relation between Initiative and Curiosity and academic outcomes compared to the other categories of ATL. Curiosity has been found to account for less than 5% of the variance in language acquisition and mathematic proficiencies in preschool whereas children's initiative has been found to account for nearly a third of the variance in preschool academic assessments (Jirout & Klahr, 2012; Dobbs, Doctoroff, Fisher, & Arnold, 2006). There are also some indications that these skills may be implicated in future outcomes. A larger body of research which links academic success to more broadly defined non-cognitive skills, that include initiative and curiosity, also provides evidence to foster these skills within early childhood intervention (Heckman, 2006). Taken together, Head Start's

emphasis on each of these categories reflects a commitment to fostering life-long learners that independently pursue and engage in classroom learning.

Assessment of Multidimensional Approaches to Learning

This body of empirical evidence on Approaches to Learning calls for Head Start to assess the multidimensional nature of this domain of skills. At the time of the 2007 Reauthorization of Head Start, the *Preschool Learning Behaviors Scale* (PLBS) was the only multidimensional assessment available for research on Approaches to Learning (McDermott, Leigh, & Perry, 2002; NRC, 2008). The PLBS is a 29-item teacher-report assessment developed in partnership with expert early childhood practitioners to be used in research and program evaluation. Psychometric review of this measure demonstrated support for validity guidelines outlined in the NRC report. It provided an initial multidimensional understanding of this construct in early childhood and had validity evidence supporting its use with Head Start students as well as evidence of temporal stability and interobserver agreement (Fantuzzo, Perry, & McDermott, 2004; McDermott et al., 2012).

Factor analyses on a nationally representative sample demonstrated that the PLBS items form three dimensions of Approaches to Learning including *Competence Motivation*, *Attentional Persistence*, and *Attitude Toward Learning*. *Competence Motivation* measures a child's propensity to engage in new tasks and concerted efforts at assigned work (McDermott et al., 2014). *Attentional Persistence* captures proficiency in sustained engagement with learning activities and children's ability to resist distractions (McDermott et al., 2014). The third dimension, *Attitude Towards Learning*, reflects children's willingness to be helped in difficult situations and cooperativeness in group

activities. Analyses using the three dimensions of the PLBS established concurrent and predictive validity to other measures of classroom behavior and cognitive functioning (McDermott et al., 2002). These dimensions captured primarily what Head Start would recognize as children's ability to self-regulate cognitively, behaviorally and emotionally in the classroom. These dimensions were not designed to measure curiosity or creativity skills. Curiosity and creativity were not categories of Approaches to Learning under Head Start's original Head Start Child Outcomes Framework published in 2000. As the knowledge base grew to recognize these important skills, the University of Pennsylvania research team sought to further develop the capacity of their teacher-report scales of ATL.

In 2011, McDermott and colleagues developed the *Learning-To-Learn Scales* (LTLS) to build upon on the original research with the PLBS. The item pool was expanded from 29 to 55 items that reflected Head Start practitioners' knowledge of their students' skills as well as a growing body of empirical literature on children's Approaches to Learning skills. A traditional bifactor model of factor analysis identified a general factor and seven dimensions of Approaches to Learning measured by the 55 items of the LTLS. The first dimension, *Effectiveness Motivation,* captures children's abilities to persevere through difficult tasks even when faced with distractions. The second dimension, *Sustained Focus in Learning,* represents children's ability to maintain attention in individual and group activities. The next three dimensions, *Demonstrated Engagement in Learning, Interpersonal Responsiveness in Learning*, and *Group Learning* capture behavioral and emotional skills that children exhibit in the context of the demands posed by a preschool classroom environment. In particular, *Demonstrated*

Engagement in Learning monitors children's ability to vocally demonstrate skills and knowledge whereas *Interpersonal Responsiveness in Learning* assesses children's restraint from aggression when frustrated and attentiveness when spoken to by the teacher. In complement, *Group Learning* measures children's capabilities to initiative activities with other children in the classroom. The last two dimensions, *Acceptance of Novelty and Risk* and *Strategic Planning,* respectively measure confident-risk taking skills in the classroom and children's abilities to creatively think through multiple solutions to a problem.

Predictive validity research demonstrates that the dimensions of the LTLS forecast a substantial reduction in the risk of future academic non-proficiency (McDermott et al., 2011). Multilevel logistic regression models showed the ability of Approaches to Learning to estimate a reduction in the likelihood of non-proficiency in cognitive functioning six months later (McDermott et al., 2011). For every area of academic testing (alphabet knowledge, vocabulary, listening comprehension, mathematics), substantial risk reduction was provided by two to four different LTLS factors (McDermott et al., 2011). The risk reduction ranged from an 87% reduction in risk for future non-proficiency in mathematics to a 44% risk reduction for non-proficiency in listening comprehension (McDermott et al., 2011). These validation analyses established the predictive validity between the multiple dimensions of the LTLS and children's outcomes.

Accounting for the Multiple Dimensions of Approaches to Learning

While the research literature connects each dimension of Approaches to Learning to various outcomes, we cannot fully understand the distinctive contribution of each one

unless we examine them simultaneously in a multidimensional context (Li-Grining, Votruba-Drzal, Maldonado-Carreño, & Haas, 2010). A simultaneous analysis of multiple dimensions allows researchers to be more precise in two important ways. First, it identifies which factors contribute relatively more to observed effects while statistically controlling for related constructs that may confound true statistical relations (Tabachnick & Fidell, 2012). Second, it increases the statistical precision of analytic models by accounting for more unique variance in important outcomes which then increases the probability of detecting important statistical relations (Cohen, 1992). Multidimensional analyses provide informative conclusions about how multiple ATL skills work in concert with one another, rather than a limited view of only individual facets of ATL in isolation.

Multivariate approaches identify which types of ATL uniquely contribute to child outcomes controlling for other types of proficiencies. In multivariate models, children's ability to engage with a task and focus despite distractions predicted preschool and elementary school outcomes while creative problem solving did not significantly predict to these outcomes (Coolahan, Fantuzzo, Mendez, & McDermott, 2000; McDermott et al., 2011; McDermott, Rikoon, Waterman, & Fantuzzo, 2012). Multivariate (i.e., canonical) correlation analyses show that attention and persistence proficiencies are the strongest predictors of peer and classroom disruptions compared to other types of ATL skills (Coolahan et al., 2000; McDermott et al., 2011). Higher scores on attention, persistence, and engagement skills contribute to the prediction of disruptive interactions, withdrawn behavior, and mathematics scores in kindergarten and first grade (Coolahan et al., 2000). This pattern of findings suggests that these attention, persistence, and engagement skills may be driving the findings that link ATL skills to early school outcomes. Children's

initiative in the classroom may contribute to outcomes when used in isolation, but these multivariate studies demonstrate that initiative does not assist the prediction of later academic outcomes above and beyond other skills (Coolahan et al., 2000; McDermott et al., 2012; McDermott et al., 2011).

Multivariate analyses of ATL also provide evidence of increased predictive validity when including multiple ATL skills in the same model. Incorporating a general factor of ATL skills, which encompassed many types of ATL proficiencies, doubled the predictive validity for academic outcomes compared with regression models that only looked at one dimension of ATL at a time (McDermott et al., 2011). A general factor of ATL increased the predicted risk-reduction of non-proficiency in preschool mathematics from 46% to 87% (McDermott et al., 2011). The general factor also increased the predicted risk-reduction of non-proficiency in listening comprehension from 44% to 81% (McDermott et al., 2011). This research illustrates how examining multiple ATL skills within multivariate models enables researchers to draw more nuanced conclusions about the unique and combined importance of ATL skills.

Common Variance among the Dimensions of the Learning-to-Learn Scales

At present the LTLS is the most highly developed multidimensional assessment of Approaches to Learning and as such it can serve as an important means to support preschool children's development. However, there are some important issues with this instrument that need to be addressed to increase its precision as a measure used in prekindergarten classrooms. Exploratory factor analysis of the LTLS items revealed high correlations between each of the seven dimensions. These high inter-factor correlations indicated that the dimensions share a significant source of common variance. The original

validation of the LTLS specified a bifactor model to account for common variance (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012).

A bifactor model contains both a general factor, that is assumed to cause common variance, and the specific dimensions that explain variance above and beyond the general factor (Chen et al., 2012). In the bifactor model, the common variance is assumed to be caused solely by the child and does not test for other relevant sources of common variance. *This bifactor model does not determine whether the common variance on an assessment actually represents a child's unique ability (i.e., a general factor of Approaches to Learning) or whether this common variance can be attributed to another source.* Factor analytic investigations of Head Start assessments often use common factor analytic approaches, like the bifactor model, but Reise and colleagues (2005) caution against using these traditional analytic techniques with nested data (e.g. students nested within teachers and classrooms).

Studies in early childhood education have convincingly demonstrated that a substantial amount of variance in a child's assessment score can typically be attributed in part to the teachers who provide the assessment ratings This variance is called *assessor variance,* which accounts for the teacher and other classroom level sources of variance. (Waterman et al., 2012). When a single teacher rates a classroom of children, it is logical to assume that the children are not the only source of assessment variance, but that teachers' subjective interpretations make up some component of the overall score variance at the classroom level (Little, 2013; Stapleton et al., 2016). This notion is confirmed by empirical research that shows not only is there 'child-variance' in teacher-report data, but there is also significant assessor variance even in highly developed

teacher-report assessments[1]. Many researchers have made the case that assessor variance attributable to the classroom teacher should not be ignored during research with teacher-report child assessment (Waterman et al., 2012).

Computer simulation research shows that ignoring even small amounts of assessor variance (i.e., 5% of the total score variance) can produce misleading results (Pornprasertmanit et al., 2014). This is because traditional statistical methods, such as the bifactor model, *assume that no assessor variance exists*. By assigning all the variance to the child, these traditional methods fail to acknowledge the common classroom context or teacher-rater as potential components of variance in the assessment. This problem worsens as assessor variance increases. In datasets with greater than 5% assessor variance, use of traditional statistical methods produces even more misleading results by failing to account for this significant source of variance in children's scores (Pornprasertmanit et al., 2014). Thus, 5% assessor variance is typically considered the threshold at which researchers should be concerned about the impact of assessor variance on the psychometric integrity of an assessment.

Studies of preschool and elementary school teacher-report assessments consistently report that the amount of assessor variance ranges from 22% to 69%, with an average of approximately 33% of the variance attributable to the teacher rater (Kim et al., 2016; Waterman et al., 2012). These findings indicate that teacher-report of young

[1] This assessor variance can be caused by a teacher-rater but can also be attributed to the aggregation of children at the classroom level and other factors (Waterman et al., 2012). Since the teacher usually rates all of the children in their classroom, it is often not possible to separate the classroom and teacher components of variance (Waterman et al., 2012). These multiple sources of variance are referred to as 'assessor variance' in early childhood assessment research, since they can be jointly recognized by identifying the assessor recording the child's proficiencies (Waterman et al., 2012).

children typically generates high levels of assessor variance that greatly exceed the 5% threshold to be concerned (Waterman et al., 2012). Teachers often must complete assessments in addition to numerous other responsibilities that leave them with little time and resources to focus on student differences when filling out assessment forms (NRC, 2008). This results in assessment data with less precisely defined differences between children assessed by the same teacher, thereby resulting in a large amount of variance associated with the teacher assessor (Waterman et al., 2012). These findings suggest that early childhood teacher-report assessments violate the assumption made by traditional analytic methods that there is *no assessor variance* present in the child assessment data.

Multilevel Statistical Methods

Complex data collected in educational settings, where students are nested within classrooms that are nested within schools, necessitates statistical considerations beyond traditional statistical approaches (Raudenbush & Bryk, 2002). The pioneering work of Raudenbush & Bryk (2002) and Goldstein (1995) demonstrated the importance of using "multilevel" statistical methods to address concerns presented by such nested data. Multilevel models, in comparison to traditional statistical approaches, afford more precision by making important distinction at these multiple 'levels' (e.g., student-level, classroom-level, and school-level). Statistical methods that ignore these distinctions among 'levels', can produce misleading results by failing to differentiate across the multiple levels (Raudenbush & Bryk, 2002). Ignoring the nested nature of the data can result in biased model estimates and standard errors, which in turn produces incorrect confidence intervals and tests of statistical significance which are key to drawing conclusions from the results (Guo & Zhao, 2000). By examining these 'levels' separately,

multilevel methods provide more detailed and precise information that otherwise would have gone undetected, since a traditional analysis an undifferentiated average of results instead of specific results at each level (Raudenbush & Bryk, 2002). Studies comparing traditional and multilevel methods for handling nested data have shown repeatedly that multilevel methods are the superior approach—so much so that the traditional ways of handling nested data have been 'discredited' (Raudenbush and Bryk, 2002; p. xx).

In the last decade, multilevel methods have been applied in psychometric sciences. One such application is *multilevel factor analysis* (Kim et al., 2016). Multilevel factor analysis is a type of multilevel model that appropriately handles nested data when testing for the presence of latent factors among assessment items. Multilevel factor analysis provides an opportunity to explore issues like assessor variance in Head Start, which results from the nested nature of teacher-report child assessment. Multilevel factor analysis is usually conducted in a multistage process (Muthen, 1993). This process involves conducting a traditional "single-level" factor analysis of the total variance (i.e., assessor variance and child variance that are undifferentiated). The next step is to conduct a multilevel factor analysis which separates out the assessor-level variance to allow for analysis of only the child-level variance (Stapleton et al., 2016), thus producing a more 'child-centered' analysis. By taking multiple stages, the results from the traditional and multilevel factor analyses facilitates a comparison between the factor structures based on the traditional total variance and the multilevel child variance. This contrast enables researchers to assess the impact of removing assessor variance. If the child-centered factors differ from the traditional factors, then researchers can conclude that removing the assessor variance meaningful consequences for the latent factors that emerged from the

assessment items.

If any differences emerge in the factor structure produced by the traditional and multilevel factor analysis, multilevel regression can then be used to explore the external validity of children's scores on the traditional versus child-centered factors. This step of the analysis determines whether there are differences in outcomes that the traditional versus child-centered factors predict in the 'real world'. Differences in predictions of 'real world' outcomes underscore the critical importance of isolating the child variance within the factor model for the ability of the assessment to provide valid results to best serve the needs of children.

Research Applications of Multilevel Methods Used in the Development and Evaluation of Assessments

A growing body of research documents the contribution of that multilevel methods make to the scientific validation of new multidimensional assessments (Kim et al., 2016). Many investigations using multilevel methodologies find that factor models change when explicitly modeling the multiple sources of variance in their data (e.g., 'child' and 'assessor'). D'Haenens and colleagues found four factors with a traditional factor analysis, but five factors emerged when multilevel factor analysis accounted for assessor variance in the model. Schweig (2014) found that an item was assigned to two factors in a traditional analysis (i.e., a double loading), but stayed on only one factor after accounting for assessor variance. In another example, Reise and colleagues (2005) found the same items loaded on each factor in the traditional versus multilevel factor models, but the strength of the item-to-factor correlations (i.e. the factor loadings) changed. Taken together, these studies illustrate that accounting for assessor variance can produce more

precise factor analytic results for factor analysis in many different ways, including the number of factors, the items assigned to each factor, or the factor loadings themselves.

These important differences in the factor structure of an assessment is not merely a technical issue because it may even mean the differences between supporting or overturning important policy decisions (Guo & Zhao, 2000). In an infamous example, Bennett (1976) showed that Great Britain elementary school students benefitted from a formal style of teaching using traditional statistical techniques, but this was overturned by Aitkin et al. (1981) who found that Bennett's result was no longer significant once multilevel analyses were conducted. Similar examples are found in the burgeoning multilevel factor analysis literature. Schweig (2014) found that a traditional analysis suggested that a school's "Distributed Leadership" score would not be an important predictor of planned teacher departure (Schweig, 2014, p. 276). In comparison, after accounting for assessor variance, Distributed Leadership did significantly predict planned teacher departure (Schweig, 2014). This would suggest that researchers should explore the relation between a school's Distributed Leadership and planned teacher departure. This critical difference in the pattern of predictions demonstrates that not accounting for assessor variance can ultimately misdirect inferences for policy which illustrates the importance of multilevel analyses to support important educational decisions.

Multilevel Analyses of Teacher-Report Child Assessments

Very few studies have applied multilevel factor analysis to teacher-report child assessment. However, two recent studies documented a difference in the number of factors extracted and conceptual reinterpretations of the factors that emerged when comparing a traditional and multilevel factor analytic approach (Peters, Algina, Smith &

Daunic, 2012; Barghaus, LeBoeuf, Fantuzzo, Brumley, & Coe, 2017). Peters et al. (2012) compared the use of multilevel to traditional factor analysis for teacher-report of elementary school students' executive functioning. The multilevel factor analysis extracted more factors than the traditional factor analysis, thereby providing greater differentiation of children's executive functioning skills. Barghaus, LeBoeuf, Fantuzzo, Brumley, & Coe (2017a) found that only two factors emerged instead of the hypothesized three factors when they applied multilevel factor analysis to a teacher report assessment of kindergarten children's engagement behaviors. The hypothesized general factor of engagement, previously found in a traditional factor analysis, did not emerge in a multilevel analysis, suggesting that a multilevel analysis may account for common variance among a set of items rather than a general factor of engagement. Such profound differences between factor structures produced by traditional versus multilevel methods underscores the importance of considering multilevel factor analysis when developing and validating teacher-report child assessments.

Similarly, taking assessor variance into account using multilevel regression methods has been shown to increase the precision of external validity analyses. Two studies using multilevel factor analysis to improve teacher-report assessment of children in kindergarten found that the correlations between multilevel factors and later academic outcomes were stronger than correlations between the traditional factors and academic outcomes. Howard and colleagues (2016) found that multilevel methods provided a 25% improvement in the strength of the association between school readiness in kindergarten and later academic outcomes. As such, their findings illustrate a small improvement in external validity when using multilevel methods (Cohen, 1992). Barghaus et al. (2017)

found a more severe difference in that the multilevel factors of kindergarten classroom engagement explained an average of three times more variance in later academic outcomes compared to the traditional factors. This meant that a traditional analysis concluded a small effect size relation between engagement and academic outcomes. In comparison, after controlling for the nested data with a multilevel analysis, the authors would have concluded a stronger medium effect size relation between engagement and academics. The changes to external validity evidence illustrate the important of multilevel methods over traditional statistical methods when analyzing teacher-report child assessment data.

Purpose of this Study

This study is motivated by the national need for a more scientifically precise early childhood assessment of the Approaches to Learning that can be used by teachers in Head Start. Currently the Learning-to-Learn Scales is the most advanced assessment that has been validated for Head Start preschool children. High LTLS's inter-dimensional correlations indicate the need to critically examine the influence that distinctive teacher variance plays in the determination of the existing dimensions of the LTLS and their external validity. To date, teacher assessor variance has not been identified and removed if necessary in the investigation of LTLS internal and external validity evidence.

The present study will test the primary hypotheses that emerge from the multilevel factor analysis research literature of teacher-report, multidimensional assessments. Three sequential research hypotheses will be tested to differentiate between the teacher- and child-levels of variance in teacher reported LTLS assessments by employing multilevel regression and factor analyses that looks across these multiple

levels (child and assessor).

- **Research Hypothesis 1:** The LTLS, as a teacher-report measure of Approaches to Learning, has items that contain a significant amount of assessor variance warranting the use of multilevel factor analytic methods
- **Research Hypothesis 2:** The use of the more empirically defensible multilevel factor analytic method will produce a distinctively different latent factor structure of the LTLS than one produced by a traditional factor analytic method
- **Research Hypothesis 3:** The factor structure of the LTLS resulting from multilevel factor analysis will result in significant differences in the external validity of LTLS factors compared to those resulting from traditional factor analysis, evidencing one or more factors with external validity to cognitive school readiness domains

The primary aim of this research is to seek a more precise *child-centered* determination of the validity of the LTLS. It is hypothesized that by attending to teacher- and child-level variance the present research study will increase the scientific accuracy and precision of this well-developed scale and thereby enhance the actionable intelligence that this assessment instrument can contribute to improve the effectiveness of Head Start intervention.

CHAPTER 2: METHODOLOGY

The current study involved a secondary analysis of data from a randomized evaluation of the efficacy of a comprehensive early childhood intervention– the *Evidence-based Program for the Integration of Curricula* (*EPIC*) (Fantuzzo, Gadsden, & McDermott, 2011). The EPIC team developed the intervention for children living in low-income households that attended Head Start centers administered by the School District of Philadelphia. The Interagency School Readiness Consortium provided support for both the development and implementation of the EPIC project. This consortium included the Administration for Children and Families, the Assistant Secretary for Planning and Evaluation, the U.S. Department of Education: Office of Special Education Programs, the Institute for Educational Sciences, and the National Institute of Child Health and Human Development. The EPIC developers created and validated the Learning-to-Learn Scales (LTLS), a teacher-report measure of Approaches to Learning, to assess changes in children's learning-related skills over the course of the intervention. The current study used data from the EPIC project to examine the presence and impact of assessor variance among the LTLS response items. The following sections provide information on the sample, measures, study design, missing data and data analytic plan.

Sample

The current study analyzes a subset of data from the larger EPIC Project encompassing 2,631 student participants across 80 Head Start classrooms. The current analytic sample included EPIC participants with baseline LTLS assessment data (see following sections on study design and missing data). This sample contained information on 2,027 unique children across 75 Head Start classrooms. Participating students ranged

in age from 35 to 69 months (M = 43.3, SD = 6.8), 51% were girls, 74% were Black/African American, 14% Hispanic/Latino, 5% White/Caucasian, and 7% mixed-race or other minorities. Approximately 12% of the sample identified as dual-language learners and 10% demonstrated special needs. The 75 teachers in the study had 2 to 44 years of teaching experience (M = 15.7, SD = 10.2), most of this being in a Head Start setting (M = 9.7, SD = 8.3).

Measures

Learning-to-Learn Scales. The Learning-to-Learn Scales (LTLS) is a teacher-report assessment of children's Approaches to Learning. It comprises 55-items that record teachers' ratings of children's Approaches to Learning behavior in the classroom. The LTLS items have three response choices of "Does not apply," "Sometimes applies," or "Consistently applies" that indicate the frequency of children's observed learning behavior (McDermott et al., 2011). McDermott et al. (2011) found seven specific dimensions and one general dimension of Approaches to Learning using exploratory and confirmatory factor analytic methods, described in Chapter 1: Introduction. Individual growth curve estimation showed the capacity of the LTLS scores to detect children's growth across six months of the preschool academic year. Additionally, McDermott and colleagues (2011) provided evidence to support concurrent validity based on the relations with other validated academic achievement tests. Several dimensions of the LTLS including *Vocal Engagement in Learning*, *Sustained Focus in Learning*, *Strategic Planning and Interpersonal Responsiveness in Learning* demonstrated predictive validity for future academic proficiency (McDermott et al., 2011). The dimensions of the LTLS explained nearly a quarter of academic ability in mathematics, alphabet knowledge,

vocabulary and listening comprehension (McDermott et al., 2011).

Learning Express. The Learning Express (LE) is an individually-administered, adaptive assessment of children's language and mathematics skills referenced to federal and state indicators of academic readiness (McDermott et al., 2009). The LE contains 325 items distributed over two equated forms and four subscales of academic achievement (Alphabet Knowledge, Vocabulary, Listening Comprehension, and Mathematics). The instrument captures 56 distinct subskills representing a wide range of item difficulty and breadth of coverage for preschool academic content. Adjusted basal and ceiling adaptive testing ensures the administration is limited to 30 minutes to minimize participant fatigue. The developers created a procedure for generating scores for the four subscales with two-parameter Item Response Theory models. McDermott et al. (2009) demonstrated the concurrent validity of the assessment through correlational analyses between the LE and other nationally norm-referenced academic tests of early reading, mathematics and receptive vocabulary (McDermott et al., 2009).

Study Design

The developers implemented the EPIC intervention in classrooms for two academic years (Fantuzzo et al., 2011). Teacher participants in the project reported on their students using the LTLS in December and May of each year of the EPIC intervention. Additionally, external assessors responsible for assessing the efficacy of the intervention administered the LE battery to each child four times (October, January, March and May) in each year of the project. The present study used data from the first year of children's participation in the EPIC project to avoid violations of statistical independence (Raudenbush & Bryk, 2002). Using these data, we specified the multilevel

and traditional factor models (see Analytic Strategy for Research Hypothesis 1 and 2 below) on data from December LTLS assessments. We tested concurrent and predictive validity of the LTLS (Research Hypothesis 3) using LE assessment data from the corresponding January and May academic assessments.

Data Analytic Plan

Taking a multilevel approach to factor analysis requires testing the assumptions imposed by a traditional factor analysis. Traditional factor analysis derives factor solutions from the total variance (i.e., variance that reflects both the child variance and the assessor variance). This traditional approach assumes that factor solutions identified with the total variance are identical to solutions derived from child-only variance in three ways: 1) they have the same number of factors, 2) the same items comprise these factors, and 3) the strength of the associations (i.e., loadings) between the items and the factors are the same (Meredith, 1993). Recent studies have demonstrated that partitioning out the assessor variance with a multilevel factor analysis can result in factor solutions that differ in these three ways (Stapleton et al., 2016; Schweig, 2014). If any of the three assumptions are violated, regression analyses using the different factor analytic results could also differ (Schweig, 2014). An external validity regression analysis can indicate whether any observed violations impact practical inferences relative to relations to important external criteria (Schweig, 2014).

The first two research questions investigated in this study involved testing the assumptions of traditional factor analysis by using Muthén's (1994) approach to multilevel factor analysis. This method involved first estimating the intraclass correlation coefficient (ICC) for each item to identify the amount of assessor variance associated

with the items (see Analytic Strategy for Research Hypothesis 1). Next, we carried out a traditional factor analysis on the total variance, ignoring the grouping of children within teacher assessors. As well as factor analyzing a correlation matrix based solely on the estimated child variance (i.e., the multilevel approach). We then compared the results from the multilevel factor solution with traditional factor solution to check for differences in the number of factors, the items that form these factors, and the loadings between items and the factors. Any differences between the factor solutions indicate a violation of the assumptions of traditional factor analysis and demonstrate the need for multilevel factor analysis. The third and final research question involved investigating whether a multilevel factor solution for the LTLS provides better concurrent and predictive validity for children's academic achievement compared to the factor solution derived from the traditional approach (see Analytic Strategy for Research Hypothesis 3). Details on the analytic strategies for each research question are provided below.

Analytic strategy for research hypothesis 1: The LTLS, as a teacher-report measure of Approaches to Learning, has items that contain a significant amount of assessor variance warranting the use of multilevel factor analytic methods.

We first estimated the intraclass correlation coefficient (ICC) for each item of the LTLS. ICCs represent the proportion of variance that is attributable to clustering of children's scores within teacher assessors in the data.[2] Each item ICC ranges from zero to

[2] This variance can be caused by the teacher-rater or it can be attributed to variability between the classroom contexts (Waterman et al., 2012). Since the teacher assessor rates all of the students in their classroom, it is not possible to separate the classroom and teacher components of variance (Waterman et al., 2012). As such, these multiple sources of variance are referred to as "assessor variance" in the literature since they are specified using the assessor's identifying variable (Waterman et al., 2012).

one with higher values indicating more of the variance is attributable to clustering.[3] To aid with interpretability, we multiplied each item ICC value by 100 to convert it to the percentage of each item's variance attributable to the assessor (i.e., a percentage of 'assessor variance'; Waterman et al., 2012). These assessor variance percentages range from 0% to 100%, with higher percentages indicating that more assessor variance is identifiable in the item variance. Past simulation research with multilevel factor models shows that even when ICCs are low (i.e., ICC = .05), clustering can still influence model fit indices and standardized parameter estimates thereby producing different results than found with a traditional factor analysis (Pornprasertmanit et al., 2014). As such, we identified whether the LTLS items exhibited ICC values above .05 (i.e., above 5% assessor variance) to benchmark the need for multilevel factor analysis. Items consistently demonstrating ICC values above .05 indicate that a multilevel factor model could produce different results than a traditional factor analysis.

Analytic strategy for research hypothesis 2: **The use of the more empirically defensible multilevel factor analytic method will produce a distinctively different latent factor structure of the LTLS than one produced by a traditional factor analytic method.**

Muthen's (1994) approach to multilevel factor analysis calls for the analyst to first carry out a traditional factor analysis. Conducting a traditional analysis allows for a comparison among any differences that arise between the multilevel and traditional

[3] Classically, the ICCs are calculated for continuous data using a standard linear model. We specified a generalized probit model for each of the 55 LTLS items because the LTLS item response data are ordinal. The variance components from each probit regression model are transformed to represent the proportion of variance associated with the teacher assessor (Dunn et al., 2015, Little, 2013).

approaches. To ensure a fair comparison, we determined that important factors must be held constant across the multilevel and traditional approaches including the estimator, and the factor rotation procedure (Ford, MacCallum, Tait, 1996; Osborne & Costello, 2009). Rather than using McDermott and colleagues (2011) traditional factor solution for the LTLS which would differ from any multilevel model we would estimate, we carried carry out a new traditional factor analysis using the same estimator and factor rotation procedure. This allowed us to attribute differences in the factor solutions to the traditional versus multilevel approaches rather than these other factors that could influence the final factor solution.

We estimated this traditional factor model using the Mplus version 7.2 Exploratory Factor Analysis procedure with the Oblimin factor rotation (Muthen & Muthen, 2016).[4] Velicer's Minimum Average Partial (MAP; Velicer, 1976) test suggested the number of factors that might best fit the data. MAP generates an estimate of a plausible number of factors to extract from the data. Rather than only examining this single solution, researchers extract factor solutions with a few less and a few more factors for comparison. This approach is analogous to estimating a confidence interval in addition to a point estimate. To select which of these solutions best fit the data, we examined goodness-of-fit indices and the acceptability and practical utility of the solution. Specifically, the determination of the final structure was based on the extent to which the solution satisfies the following criteria: (a) goodness-of-fit through Root-

[4] Oblimin, an oblique rotation, allows the factors to correlate and is commonly employed in multilevel factor modeling. Other rotation methods, such as Geomin, are possible in a multilevel framework, while others like Promax are not. The differential performance of these rotations have not been extensively studied in a multilevel framework.

Mean-Square Error of Approximation (RMSEA) values below .06 and Standardized Root Mean Residual (SRMR) values below .10 (Kline, 2010); (b) have at least three salient items per factor where loadings ≥ .40 indicate salience (McDermott et al., 2011); (c) produce internally consistent factors where Cronbach's $\alpha \geq .70$ indicates reliability (McDermott et al., 2011); (d) approximate simple structure as reflected in item coverage where each item loads on only one unique factor (Yates, 1987); and (e) make theoretical sense in terms of parsimonious coverage of the data and compatibility with leading research in the content area (Fabrigar, Wegener, MacCallum, & Strahan, 1999).

First, each factor solution was tested relative to its capacity to demonstrate empirical fit on both the RMSEA and SRMR. Next, we inspected each factor solution to ensure that each factor retained at least three salient items. Of these factors, we assessed each factor for adequate reliability ($\alpha \geq .70$). Any factor solution demonstrating lack of empirical fit, exhibiting factors with less than three items, or solutions with unreliable factors would not be considered as a viable factor solution. We would only consider solutions that met the above criteria. Of the viable solutions, we determined which made the most theoretical sense in terms of the extant literature on Approaches to Learning and past research specifically with the LTLS (Fabrigar, Wegener, MacCallum, & Strahan, 1999).

To identify the dimensions of the LTLS based solely on the child variance, we estimated a multilevel factor model with the Mplus version 7.2 Two-level Exploratory Factor Analysis procedure (Muthen & Muthen, 2016). This procedure allows the analyst to estimate the factor structure on only the child-level variance using the "Saturation" method (Ryu & West, 2009). The Saturation method involves specifying a perfectly

fitting factor model (i.e., fully saturated with zero degrees of freedom) at the assessor-level and allows the child-level to be freely estimated (Ryu & West, 2009). Stapleton and colleagues (2016) recommend the Saturation method when the factors conceptually originate at the child-level of analysis. We employed the MAP procedure and same model selection criteria used with the traditional factor analysis to select the multilevel factor solution.

To assess whether the dimensions derived from the traditional and multilevel factor analytic approaches depart from one another, we first identified whether the solutions had the same number of factors. If the solutions had the same number of factors, we examined whether the same exact items load onto each of the corresponding factors from the multilevel and traditional models. This indicated which of the factors (if any) are invariant to factor analytic method and where discrepancies between traditional and multilevel approaches emerged. For factors that have the exact same items, we then assessed at the strength of the association (i.e., loadings) between items and the latent factors. To do so, we compared individual item loadings across the solutions. Higher loadings indicate that a particular item better defines the factor. Additionally, factors with consistently higher loadings can produce higher reliability coefficients. Differences in factor loadings between the solutions would indicate discrepancy between the traditional and multilevel factor solutions and a violation of the assumptions imposed by a traditional factor analysis.

Analytic strategy for research hypothesis 3: The factor structure of the LTLS resulting from multilevel factor analysis will result in significant differences in the external validity of LTLS factors compared to those resulting from traditional

factor analysis, evidencing one or more factors with external validity to cognitive school readiness domains.

In the case that the dimensions derived from a multilevel factor analysis (i.e., "child-centered" dimensions) differ from the traditional analysis, it is possible that the child-centered dimensions could predict children's academic progress differently than the traditional dimensions. To test this, we calculated linear composite factor scores from the resulting multilevel dimensions and traditional dimensions and used them to predict concurrent and future academic achievement six months later. Linear composite factor scores are sums of item values for the items comprising each factor (DiStefano, Zhu, Mindrila, 2009).[5] We used these linear composite factor scores to estimate external validity models with multilevel regression models that adjust for the clustered nature of data from children within classrooms (Raudenbush and Bryk, 2002).

We identified two levels in the external validity multilevel regression model: Level-1 contained score variation between children within classrooms, and Level-2 contained score variation between classrooms. Explanatory variables of interest to the current analysis (i.e., children's scores on the LTLS dimensions) operated at the child level (Level-1). As such, a "fixed-effects" approach to multilevel regression allowed us to estimate these Level-1 relations controlling for the variation from teacher assessors at Level-2 (Allison, 2005). A fixed-effects approach removed variance in both the predictors and the outcomes that was associated with the teacher assessor (Allison, 2005).

[5] Linear composite scores reflect only factor pattern differences in the observed factor solution. This technique has been used in previous research on multilevel factor models (Schweig, 2014) and produces a conservative test of differences between multilevel and traditional methods than refined factor scoring methods since it will not incorporate and differences observed among the loadings (DiStefano, Zhu, Mindrila, 2009).

First, we ran two sets of models for all dimensions. In the first set of models, all dimensions resulting from a traditional factor analytic approach predicted academics (i.e., "Model A"). A second set of models, we included all of these traditional dimensions *plus* all of the multilevel dimensions in one model ("Model AB"). This allowed us to assess how much *additional* variance the multilevel dimensions (Model AB) explained above and beyond the traditional dimensions (Model A).

We employed this same approach for each corresponding dimension to better understand differential predictive validity among the individual dimensions. Specifically, each dimension derived from traditional factor analysis separately predicted academic outcomes (Model A). In the second set of models (Model AB), we added the corresponding multilevel dimension to the model. This sequence of modeling allowed us to evaluate the predictive capacity of each dimension in addition to the incremental predictive capacity of the multilevel dimension.

To summarize this analysis, we used the R^2 values from Model A and Model AB to calculate Cohen's f^2 an effect size measure of variance explained within a multilevel regression model framework. This metric reflects the proportion of variance uniquely accounted for by the multilevel factors (B), over and above the traditional factors (A). Specifically, we calculated Cohen's f^2 values as per Seyla et al., (2012) as,

$$f^2 = ((R^2AB - R^2A)/(1- R^2AB)),$$

where R^2A is the proportion of variance accounted for by A relative to a baseline model with only a teacher fixed-effect $R^2_{adj\text{-}null}$ and R^2AB is the proportion of variance accounted for by A and B together relative to a baseline model with only a teacher fixed effect $R^2_{adj\text{-}\mathrm{null}}$.

We evaluated these Cohen f^2 values using conventional effect size metrics (Cohen, 1992). We considered f^2 values between .02 and.15 as small effects, values between .15 and .35 as medium effects, and values above .35 as large effects (Cohen, 1992). Because f^2 values are scaled to reflect their proportion of variance explained relative to variance explained by the full model (i.e., R^2AB) they cannot be interpreted directly as a proportion of variance explained (Seyla et al., 2012). However, for values of R^2AB closer to zero, these values will closely match variance explained calculations. For illustration, a simplifying way to interpret these f^2 values would be that an f^2 of .06 means that the multilevel dimensions (Model AB) explained roughly 6% more variance than traditional dimensions (Model A).

This Model A and Model AB approach was used for all four academic outcomes of the Learning Express (Vocabulary, Mathematics, Listening Comprehension, Alphabet Knowledge) for both concurrent and predictive validity six months later. This meant we estimated a total of 16 models for the combined dimensions for four outcomes measured at both baseline and six months later (8 outcomes for Model A and 8 outcomes for Model AB), and 16 models for each corresponding individual dimension that exhibited violations of the traditional factor analytic assumptions. This included 8 outcomes for Model A and 8 outcomes for Model AB. Each of the 16 models were necessary to perform the variance explained calculations (R^2AB and R^2A) and resulting Cohen's f^2 .

CHAPTER 3: RESULTS

Analytic Results for Research Hypothesis 1: The LTLS, as a Teacher-Report Measure of Approaches to Learning, has Items that Contain a Significant Amount of Assessor Variance Warranting the use of Multilevel Factor Analytic Methods

We first estimated assessor variance for the 55 items administered on the LTLS. [6] This involved calculating the intraclass correlation coefficient (ICC) for each item and then multiplying the ICCs by 100 so that the estimates could be interpreted as a percentage of variance. This analysis identified an average of 21% assessor variance in the items from the LTLS. As such, 21% of the variance across the items could be associated with the assessor administering the assessment. For all items, the assessor variance calculated exceeded 5% of total variability indicating that multilevel methods could produce different results than a traditional analysis (Pornprasertmanit et al., 2014). Items that captured children's sustained focus in the classroom had the highest estimates of assessor variance. For example, we could associate 37% of the variance in the item "Focused on individual activity 30 minutes" with the teacher assessor. Items associated with the least amount of assessor variance included group and peer learning skills in the classroom like "Helps, shares and discusses in group," "Maintains essential role in small group," "Seeks answers by engaging with materials and people". These items evidenced 12% to 14% assessor variance. Items with a higher amount of assessor variance indicates

[6] Assessor variance can be caused by the teacher-rater or it can be attributed to the classroom context (Waterman et al., 2012). Since the teacher usually rates all of the children in their classroom, it is not possible to separate the classroom and teacher components of variance (Waterman et al., 2012). As such, these multiple sources of variance are referred to as "assessor variance" in the literature (Waterman et al., 2012).

that teachers tend to rate the students in their classroom more similarly to each other on those items. As such, there is a lower proportion of variability between children in a classroom on items with higher amounts of assessor variance.

Analytic Results for Research Hypothesis 2: **The use of the more Empirically Defensible Multilevel Factor Analytic Method will Produce a Distinctively Different Latent Factor Structure of the LTLS than one Produced by a Traditional Factor Analytic Method**

We next examined the underlying dimensions of the LTLS using traditional factor analytic methods. Velicer's Minimum Average Partial (MAP) test indicated that six factors might best fit the data so we compared solutions ranging from five to nine factors. To select the optimal model of these five solutions, we first considered goodness-of-fit indices which estimate the fit of the model relative to the sample data. As per our criteria, we determined acceptable model fit as RMSEA and SRMR indices below .06 and .10 respectively (Kline, 2010; Dunn et al., 2015). The RMSEA indices results ranged from .05 for the 5-Factor solution to .03 for the 9-Factor solution. In addition, the 5-Factor through 9-Factor solutions all exhibited SRMR values of .02. Both the RMSEA and SRMR values indicated acceptable model fit below the criteria threshold for all five of the factor solutions.

We then assessed whether these factor solutions retained at least three items per factor and if these factors produced reliable information. We found that only three of the five factor solutions met this requirement. The 5-Factor, 6-Factor and 7-Factor solutions included at least three salient items per factor and demonstrated adequate internal consistency ($\alpha > .70$). At least one factor in each the 8-Factor and 9-Factor solution did

not contain three items (see Table 1). Because of this, we no longer considered these two solutions for the final selection.

TABLE 1

Factor Selection Criteria by Factor Solution using Traditional Factor Analysis

Factor solution	RMSEA < .06	Within SRMR < .10	At least 3 salient items (≥ |.395|) per factor	Internally Consistent (rs > .70)
5-Factor	0.05	0.02	Yes	Yes
6-Factor	0.04	0.02	Yes	Yes
7-Factor	0.04	0.02	Yes	Yes
8-Factor	0.04	0.02	No	-
9-Factor	0.03	0.02	No	-

Note. Root-Mean-Square Error of Approximation is abbreviated as RMSEA and Standardized Root Mean Residual is abbreviated as SRMR.

The remaining three factor solutions (i.e., the 5-, 6-, or 7-Factor) met our next criterion that the factor solution should exhibit good item coverage and lack of double loading items. All three evidenced a clear majority of the items (40 to 50 items) exhibiting only one, unique loading on a single factor. This was an important facet of these factor solutions because an overarching goal in factor analysis is to identify a 'simple factor structure' (Kaiser, 1974). A simple factor structure is a factor solution where the items exhibit only one unique loading on a single factor and does not have an excessive number of items that load onto more than one factor (i.e. no 'double-loaders'). Among our three solutions, the item pool was lowered (40 to 50 down from a total of 55 items) more often because all factor loadings were below .40 for a particular item (i.e., a

'non-salient loader' item) versus instances of an item saliently loading on more than one factor (i.e., a 'double-loader' item; see Table 2). The absence of double-loading items, along with the presence of so many items that load saliently on only one individual factor, indicates relatively simple factor patterns for all three viable factor solutions.

TABLE 2

Characteristics of Remaining Factor Solutions using Traditional Factor Analysis

Factor solution	Number of Items Retained	Number of Double Loaders	Number of Non-Salient Loaders
5-Factor	50	2	3
6-Factor	44	1	10
7-Factor	40	1	14

Note . Number of Items Retained are the count of items with only one salient loading. Number of Double Loaders are the number of items that saliently load on more than one factor. Number of Non-Salient Loaders are the number of items that do not saliently load on any factor. These three columns should sum to 55 total items per each solution.

Final traditional factor analytic model selection. We selected a final model that provided parsimonious coverage of the data and was consistent with theory and extant research (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Of the three solutions that evidenced model fit, we rejected the 5-Factor solution as an option for the final model. The third factor on the 5-Factor solution was conceptually unclear as it combined children's abilities to independently plan and their engagement in the classroom environment (see the 5-Factor solution in Appendix A). These two skills are viewed as correlated, but distinct, skills in the extant literature (Bustamante, White, & Greenfield, 2017; Hyson, 2008).

The 6-Factor and 7-Factor solution evidenced clearer differentiation between the

factors. However, the 7-Factor solution included children's group learning skills as a separate factor whereas the 6-Factor solution did not include this factor. On the 6-Factor solution, these group learning items were split between a factor representing strategic planning skills, and a factor indicating interpersonal skills in the classroom (see the 6-Factor solution in Appendix B). Given that children's group learning skills are a distinct facet of Approaches to Learning that corresponds to Emotional and Behavioral Self-Regulation skills (Administration for Children and Families, 2015a; Hyson, 2008), we selected the 7-Factor solution as the best fit to the data.

The 7-Factor solution replicated dimensionality from the validation of the original LTLS assessment (see Chapter 1: Introduction; McDermott et al., 2011)[7]. These seven factors represented distinct types of Approaches to Learning skills (e.g., self-regulation, curiosity, initiative) which have a research base backing their differential validity evidence. These factors encompassed *Strategic Planning* (Factor 1), *Interpersonal Responsiveness in Learning* (Factor 2), *Acceptance of Novelty and Risk* (Factor 3), *Sustained Focus in Learning* (Factor 4), *Effectiveness Motivation* (Factor 5), *Demonstrated Engagement in Learning* (Factor 6), and *Group Learning* (Factor 7). These seven factors collectively included both self-regulations skills and skills capturing creativity, curiosity and initiative skills in the classroom. See Table 3 for the factor pattern loadings of the 7-Factor traditional solution.

---

[7] To ensure a fair comparison, we determined that the factor rotation procedure must be held constant across the multilevel and traditional approaches (Ford, MacCallum, Tait, 1996; Osborne & Costello, 2009). Rather than using McDermott and colleagues (2011) traditional factor solution which would differ from any multilevel model we would estimate, we carried out a new traditional factor analysis using the same factor rotation procedure (GEOMIN). This allowed us to attribute differences in the factor solutions to the traditional versus multilevel approaches rather than these other factors that could influence the final factor solution. Differences between the results of these two methods are described in Appendix C.

The 7-Factor solution represented factors that were consistent with the federal school readiness framework for Head Start and empirical research linking these skills to success in the early classroom (see Administration for Children and Families, 2015a; Hyson, 2008). *Strategic Planning* (Factor 1) most closely aligned with Head Start's subdomain of Cognitive Self-Regulation as it captured children ability to demonstrate flexibility in thinking and behavior (ACF, 2015a, Goal P-ATL 9). *Interpersonal Responsiveness in Learning* (Factor 2) corresponds to the Emotional & Behavioral Self-Regulation sub-domain (ACF, 2015a, Goal P-ATL 1-4). These dimensions monitor the behavioral demands of responding to classroom routines and interacting appropriately with peers and adults. *Acceptance of Novelty and Risk* (Factor 3) corresponds most closely with the sub-domain of Initiative and Curiosity (ACF, 2015a, Goal P-ATL 10-11). Children developing these skills show an interest and curiosity in their classroom environment. *Sustained Focus in Learning* (Factor 4) most closely aligned with Head Start's subdomain of Cognitive Self-Regulation since it required children to be able to persist in tasks and maintain focus and attention with minimal adult support (ACF, 2015a, Goal P-ATL 6-7). They capture children's flexibility in thinking and ability to control cognitive thought processes. *Effectiveness Motivation* (Factor 5) correspond to Head Start's subdomain of Cognitive Self-Regulation. They capture children's flexibility in thinking and ability to control cognitive thought processes. *Demonstrated Engagement in Learning* (Factor 6) most closely correspond to Head Start's Early Learning Outcomes Framework conceptualization of children's *Initiative & Curiosity*. Children developing these skills show an initiative to engage in their classroom environment. Finally, *Group Learning* (Factor 7) mirrored skills under Head Start's Emotional and Behavioral Self-

Regulation subdomain of Approaches to Learning. Specifically, under Goal P-ATL 4 Head Start children are expected to be able to wait for their turn, refrain from aggressive behavior towards other, and began to understand the consequences of behavior. This 7-factor solution best reflected the Head Start framework and existing research on distinct aspects of Approaches to Learning that are predictive of children's academic outcomes.

TABLE 3

Rotated Factor Loadings for the 7-Factor Solution using the Traditional Approach

	Factor						
Item	1	2	3	4	5	6	7
Basic understanding of cause and effect	**.64**	.18	.00	.08	.06	.04	.07
Develops plan after considering consequences	**.62**	.18	.06	.08	.12	.08	-.01
Compares new task to previous re: what worked	**.56**	.08	.01	.00	.14	.15	.17
Changes strategies when solution not working	**.53**	.07	.14	.06	.24	.08	.02
Develops plan for multi-step activity	**.49**	.00	-.07	.16	.06	.23	.24
Verbalizes possible consequences	**.47**	.08	-.05	.08	.05	.29	.22
Self-corrects errors	**.47**	.11	.14	.06	.27	.04	.05
Communicates problems may have more than one solution	**.41**	.05	-.03	-.02	.28	.23	.24
Refrains from aggression when frustrated	.11	**.73**	.14	.06	-.13	-.11	-.02
Attentive when spoken to by teacher	.08	**.64**	.04	.15	.07	.11	-.02
Accepts teacher advice by following it	-.11	**.64**	.05	.15	.16	.19	.01
Listens and waits for turn to speak	.11	**.63**	-.03	.25	.04	-.05	.02
Accepts peer advice by following it	.05	**.62**	-.02	-.07	.12	-.02	.29
Responds positively to suggestions for alternate approach	.15	**.60**	.22	-.02	.00	.11	.00
Attentive when teacher leads group activity	.03	**.60**	.00	.26	.16	.06	-.04
Takes turn in group without reminder	.17	**.52**	.06	.18	.09	-.21	.23
Responds to questions about ideas without becoming upset	.20	**.49**	.25	-.12	-.03	.16	.12
Responds positively to assistance	-.13	**.42**	.04	-.04	.17	.29	.23

Note . Strategic Planning (Factor 1), Interpersonal Responsiveness in Learning (Factor 2), Acceptance of Novelty and Risk (Factor 3), Sustained Focus in Learning (Factor 4), Effectiveness Motivation (Factor 5), Demonstrated Engagement in Learning (Factor 6), and Group Learning (Factor 7).

Rotated Factor Loadings for the 7-Factor Solution using the Traditional Approach

	Factor						
Item	1	2	3	4	5	6	7
Willingly participates in unfamiliar activities	.04	.09	**.65**	-.02	.12	.07	.13
Receptive when asked to participate in new task	.02	.12	**.63**	.04	.11	.08	.10
Participates in activity or lesson	-.11	.25	**.55**	.27	.01	.10	.05
Previous attempts unsuccessful, still tries	.20	-.01	**.47**	.07	.36	.09	-.09
Shows interest and positive attitude toward new activities	-.14	.21	**.40**	.00	.27	.28	.11
Focused on individual activity, 20 minutes	.08	.00	-.05	**.79**	.13	.06	.06
Focused on individual activity, 10 minutes	-.03	.15	.13	**.72**	.01	.10	.00
Focused on individual activity, 30 minutes	.09	.00	-.11	**.70**	.17	.04	.09
Focused on group activity, 10 minutes	-.04	.20	.18	**.63**	.01	.13	.01
Self-selects activity without direction	.15	.18	.29	**.41**	-.05	-.09	.17
Tries activity when solution not forthcoming	.08	.05	.07	.12	**.70**	.03	.07
Perseveres when distracting activities available	.07	.11	.02	.20	**.70**	-.03	.06
Engages in activity previously challenging	.16	.02	.20	-.04	**.60**	.10	.06
Perseveres with little input from teacher	.18	.03	.12	.18	**.54**	-.07	.15
Practices activity without prompting	.16	.01	.17	.21	**.46**	.06	.06
Demonstrates pride in work products	.09	-.04	.18	.14	-.05	**.77**	.00
Verbalizes frustration and asks for help	-.01	.16	-.17	.06	.07	**.61**	.12
Voluntarily demonstrates academic skills	.31	-.08	.20	.13	-.09	**.60**	-.02

Note . Strategic Planning (Factor 1), Interpersonal Responsiveness in Learning (Factor 2), Acceptance of Novelty and Risk (Factor 3), Sustained Focus in Learning (Factor 4), Effectiveness Motivation (Factor 5), Demonstrated Engagement in Learning (Factor 6), and Group Learning (Factor 7).

Rotated Factor Loadings for the 7-Factor Solution using the Traditional Approach

	Factor						
Item	1	2	3	4	5	6	7
Maintains essential role in small group	.29	-.09	.08	.13	.16	.05	**.53**
Works cooperatively to complete task	-.09	.30	.13	.26	.08	-.01	**.46**
Teaches another child a skill	.36	-.03	.03	.15	.12	.09	**.45**
Helps, shares, discusses in group	.02	.24	.15	.11	.10	.15	**.43**
Asks questions and shares ideas	.20	.00	.19	.04	.04	.33	.38
Plays with child during free play	.01	.15	.26	.30	-.18	.12	.37
Seeks answers by engaging with materials and people	.15	.11	.05	.09	.15	.33	.28
Initiates activity with children	.13	.10	.28	.35	-.07	.04	.27
Guesses even when unsure	.05	.02	.27	-.06	.15	.38	.27
Asks teacher for a task	.02	.13	-.08	.01	.13	.36	.23
Identifies alternate uses for object	.26	-.04	.10	.21	.06	.27	.18
Works independently with minimal supervision	.16	.16	.09	.37	.27	-.10	.17
Engages in activity without need for approval	-.01	.04	.34	.17	.37	.07	.16
Sense of humor with errors	.26	.21	.12	-.18	.14	.29	.15
Learns by accepting constructive feedback	.05	.37	.10	.09	.22	.27	.06
Tries new task instead of familiar	.27	-.02	.35	.03	.30	.11	.01
Perseveres with assistance and encouragement	.25	.03	.35	.19	.35	.02	-.09
Screens out noise and distractions	.30	.26	.06	.31	.29	-.08	-.13
Verbalizes frustration but continues working	.05	.19	-.11	-.02	.42	.46	.09

Note . Strategic Planning (Factor 1), Interpersonal Responsiveness in Learning (Factor 2), Acceptance of Novelty and Risk (Factor 3), Sustained Focus in Learning (Factor 4), Effectiveness Motivation (Factor 5), Demonstrated Engagement in Learning (Factor 6), and Group Learning (Factor 7).

We next identified the dimensions of the LTLS assessment based solely on the child variance using a multilevel factor analytic approach. Velicer's Minimum Average Partial test indicated that seven factors might best fit the data. As per our methodology, we extracted factor solutions above and below this estimate. In total, we compared a 5-Factor Solution though a 9-Factor Solution. To select the optimal model of these five solutions, we first considered goodness-of-fit indices which estimate the fit of the model relative to the sample data. All five factor solutions exhibited RMSEA values of .03 and SRMR indices of .02. These goodness of fit indices suggested that all of the factor solutions had acceptable discrepancy between the fit of the sample data and the model (RMSEA and SRMR indices below .06 and .10, respectively).

Of the five solutions, the 6-Factor, 7-Factor and 8-Factor solutions included at least three salient items per factor and were internally consistent on their respective factor solution. The 5-Factor and 9-Factor solutions did not include at least three salient items per factor. This meant that the 5-Factor and 9-Factor solutions were no longer acceptable given our criteria. See Table 4 for a summary of the fit criteria for each respective factor solution.

TABLE 4

Factor Selection Criteria by Factor Solution using Multilevel Factor Analysis

Factor solution	RMSEA < .06	Within SRMR < .10	At least 3 salient items (≥ |.395|) per factor	Internally Consistent (rs > .70)
5-Factor	0.03	0.02	No	-
6-Factor	0.03	0.02	Yes	Yes
7-Factor	0.03	0.02	Yes	Yes
8-Factor	0.03	0.02	Yes	Yes
9-Factor	0.03	0.02	No	-

Note. Root-Mean-Square Error of Approximation is abbreviated as RMSEA and Standardized Root Mean Residual is abbreviated as SRMR.

All three of the remaining factor solutions evidenced good item coverage and lack of double-loading items. The 6-Factor and 7-Factor solutions retained 40 items out of the 55 LTLS items and 37 items were retained for the 5-Factor solution. Items were generally excluded from these solutions because they had non-salient loadings. These solutions rarely contained items that loaded on more than one factor (i.e., double-loading items). The 6-Factor solution had 16 non-saliently loading items and two double-loading items. The 7-Factor solution 15 non-salient loaders and no double-loading items. The 8-Factor solution had 14 items with non-salient loadings and one double-loading item. The absence of double loaders indicated relatively simple factor solutions (i.e., where one item uniquely loads on one single factor) for all three viable factor solutions. Table 5 presents the comparisons of characteristics among these factor solutions.

TABLE 5

Characteristics of Remaining Factor Solutions using Multilevel Factor Analysis

Factor solution	Number of Items Retained	Number of Double Loaders	Number of Non-Salient Loaders
5-Factor	37	2	16
6-Factor	40	0	15
7-Factor	40	1	14

Note . Number of Items Retained are the count of items with only one salient loading. Number of Double Loaders are the number of items that saliently load on more than one factor. Number of Non-Salient Loaders are the number of items that do not saliently load on any factor. These three columns should sum to 55 total items per each solution.

Final multilevel factor analytic model selection. Given the current criteria, three multilevel factor solutions provided a viable fit to the data (i.e., the 6-Factor, 7-Factor or 8-Factor solutions). These solutions met our goodness-of-fit criteria and all produced internally consistent factors. All three of these factor solutions very closely aligned with the seven original dimensions of the LTLS. However, both the 7-Factor and 8-Factor solution also contained a factor not found in the original validation analyses (McDermott et al., 2011). This factor was comprised the following four items: “Previous attempts unsuccessful, still tries”, “Develops plan after considering consequences”, “Basic understanding of cause and effect”, and “Perseveres with assistance and encouragement”. This factor reflected children’s consequential thinking and planning behaviors, which was interpreted conceptually identical to another factor representing strategic planning skills. Because of this conceptual overlap, we ruled out the 7-Factor and 8-Factor solutions as viable factor models.

We selected the 6-Factor multilevel solution as the optimal fit to the data. Each of these six factors had clear and unique conceptual meaning consistent with theory and

extant research. Like the traditional analysis, these six factors represented self-regulations skills, and skills measuring creativity, curiosity and initiative in the classroom. The factors were *Strategic Planning* (Factor 1), *Interpersonal Responsiveness in Learning* (Factor 2), *Acceptance of Novelty and Risk* (Factor 3), *Sustained Focus in Learning* (Factor 4), *Demonstrated Engagement in Learning* (Factor 5), and *Effectiveness Motivation* (Factor 6).

Strategic Planning (Factor 1) most closely aligned with Head Start's subdomain of Cognitive Self-Regulation as it captured children ability to demonstrate flexibility in thinking and behavior (ACF, 2015a, Goal P-ATL 9). *Interpersonal Responsiveness in Learning* (Factor 2) corresponds to the Emotional & Behavioral Self-Regulation sub-domain (ACF, 2015a, Goal P-ATL 1-4). These dimensions monitor the behavioral demands of responding to classroom routines and interacting appropriately with peers and adults. *Acceptance of Novelty and Risk* (Factor 3) corresponds most closely with the sub-domain of Initiative and Curiosity (ACF, 2015a, Goal P-ATL 10-11). Children developing these skills show an interest and curiosity in their classroom environment. *Sustained Focus in Learning* (Factor 4) most closely aligned with Head Start's subdomain of Cognitive Self-Regulation since it required children to be able to persist in tasks and maintain focus and attention with minimal adult support (ACF, 2015a, Goal P-ATL 6-7). They capture children's flexibility in thinking and ability to control cognitive thought processes. *Demonstrated Engagement in Learning* (Factor 5) most closely correspond to Head Start's Early Learning Outcomes Framework conceptualization of children's *Initiative & Curiosity*. Children developing these skills show an initiative to engage in their classroom environment. *Effectiveness Motivation* (Factor 6) corresponds

to Head Start's subdomain of Cognitive Self-Regulation. It captures children's flexibility in thinking and ability to control cognitive thought processes. The full 6-Factor solution is presented in Table 6.

TABLE 6

Rotated Factor Loadings for the 6-Factor Solution using the Multilevel Approach

	Factor					
Item	1	2	3	4	5	6
Basic understanding of cause and effect	**.53**	.28	-.06	.04	.17	.20
Develops plan after considering consequences	**.46**	.22	.10	.04	.12	.24
Compares new task to previous re: what worked	**.40**	.08	.12	.19	.27	.09
Refrains from aggression when frustrated	-.02	**.81**	.11	-.01	-.11	-.07
Accepts peer advice by following it	.11	**.72**	.02	.00	.10	.00
Listens and waits for turn to speak	.09	**.72**	-.07	.14	-.04	.08
Takes turn in group without reminder	.18	**.69**	.01	.16	-.07	.05
Attentive when spoken to by teacher	.04	**.62**	.08	.10	.10	.08
Responds positively to suggestions for alternate approach	-.02	**.62**	.29	-.06	.10	.02
Accepts teacher advice by following it	-.19	**.62**	.09	.12	.20	.17
Attentive when teacher leads group activity	.00	**.58**	-.02	.23	.06	.18
Responds to questions about ideas without becoming upset	.12	**.44**	.32	-.03	.22	-.10
Learns by accepting constructive feedback	-.01	**.42**	.07	.04	.37	.18

Note . Strategic Planning (Factor 1), Interpersonal Responsiveness in Learning (Factor 2), Acceptance of Novelty and Risk (Factor 3), Sustained Focus in Learning (Factor 4), Demonstrated Engagement in Learning (Factor 5), and Effectiveness Motivation (Factor 6).

Rotated Factor Loadings for the 6-Factor Solution using the Multilevel Approach

	Factor					
Item	1	2	3	4	5	6
Willingly participates in unfamiliar activities	.03	.01	**.79**	.08	.03	.04
Receptive when asked to participate in new task	.01	.06	**.75**	.05	.05	.05
Participates in activity or lesson	-.03	.26	**.62**	.16	.01	.05
Shows interest and positive attitude toward new activities	-.13	.14	**.52**	.03	.28	.18
Focused on individual activity, 10 minutes	-.12	.17	.09	**.69**	.06	.13
Focused on individual activity, 20 minutes	.09	.13	.04	**.68**	-.04	.16
Focused on individual activity, 30 minutes	.10	.04	.03	**.67**	-.03	.19
Focused on group activity, 10 minutes	.00	.23	.20	**.50**	.06	.09
Maintains essential role in small group	.31	-.06	.15	**.41**	.32	-.01
Demonstrates pride in work products	.04	.00	.21	-.04	**.71**	.07
Verbalizes frustration and asks for help	-.09	.13	-.07	.03	**.71**	.00
Guesses even when unsure	.06	-.06	.26	.05	**.57**	.11
Verbalizes frustration but continues working	-.01	.19	-.07	.01	**.57**	.32
Voluntarily demonstrates academic skills	.24	-.01	.23	-.09	**.55**	.09
Verbalizes possible consequences	.31	.06	-.05	.10	**.54**	.16
Seeks answers by engaging with materials and people	.12	.08	.08	.21	**.53**	.07
Asks questions and shares ideas	.24	-.02	.22	.20	**.52**	-.06
Asks teacher for a task	.04	.16	-.01	.11	**.49**	-.03
Communicates problems may have more than one solution	.27	-.02	.08	.20	**.43**	.19

Note . Strategic Planning (Factor 1), Interpersonal Responsiveness in Learning (Factor 2), Acceptance of Novelty and Risk (Factor 3), Sustained Focus in Learning (Factor 4), Demonstrated Engagement in Learning (Factor 5), and Effectiveness Motivation (Factor 6).

Rotated Factor Loadings for the 6-Factor Solution using the Multilevel Approach

	Factor					
Item	1	2	3	4	5	6
Tries activity when solution not forthcoming	.00	.04	.11	.21	.12	**.64**
Perseveres when distracting activities available	.03	.12	.03	.26	.08	**.63**
Engages in activity previously challenging	.12	.03	.18	.05	.22	**.52**
Practices activity without prompting	.11	.06	.15	.19	.15	**.47**
Perseveres with little input from teacher	.23	.06	.17	.25	.00	**.43**
Screens out noise and distractions	.25	.35	-.05	.19	-.07	.37
Tries new task instead of familiar	.26	.08	.31	-.04	.11	.35
Self-corrects errors	.37	.13	.21	.10	.06	.30
Changes strategies when solution not working	.38	.07	.17	.13	.13	.26
Helps, shares, discusses in group	.10	.28	.15	.24	.35	.00
Responds positively to assistance	-.14	.43	.12	-.01	.44	.07
Works cooperatively to complete task	.01	.37	.19	.39	.15	-.06
Previous attempts unsuccessful, still tries	.23	.06	.46	-.09	-.03	.44
Perseveres with assistance and encouragement	.24	.10	.36	.06	-.06	.39
Engages in activity without need for approval	-.02	.03	.33	.24	.18	.32
Teaches another child a skill	.34	.00	.09	.36	.34	.02
Works independently with minimal supervision	.20	.23	.06	.36	.02	.26
Self-selects activity without direction	.15	.25	.22	.34	.03	.03
Initiates activity with children	.24	.20	.28	.33	.11	-.09
Identifies alternate uses for object	.19	-.01	.12	.31	.35	.03
Develops plan for multi-step activity	.35	-.03	.03	.29	.39	.09
Plays with child during free play	.18	.29	.27	.29	.21	-.26
Sense of humor with errors	.22	.20	.22	-.10	.37	.05

Note . Strategic Planning (Factor 1), Interpersonal Responsiveness in Learning (Factor 2), Acceptance of Novelty and Risk (Factor 3), Sustained Focus in Learning (Factor 4), Demonstrated Engagement in Learning (Factor 5), and Effectiveness Motivation (Factor 6).

Our next task was to assess whether the traditional approach and multilevel approach produce factor solutions with the number of factors; that the same items that comprise these factors; and finally, that the strength of the associations between the items and the factors (i.e., item loadings) remain the same (Meredith, 1993). These three components are broken down into the following sections listed below.

The number of factors. We selected a 7-Factor solution as the optimal model for the traditional analysis and a 6-Factor solution for the multilevel analysis. Six of these dimensions emerged in both analyses, namely *Strategic Planning*, *Demonstrated Engagement in Learning*, *Sustained Focus in Learning*, *Acceptance of Novelty and Risk*, *Effectiveness Motivation*, and *Interpersonal Responsiveness in Learning*. The seventh factor identified using the traditional factor analytic approach, *Group Learning,* did not emerge in the multilevel solution. Three of the four items that comprised the traditional *Group Learning* factor failed to load onto any of the six factors in the multilevel factor solution. These three items did not emerge on any other factor in the multilevel analysis. This indicates that the *Group Learning* items did not represent their own distinct dimension, nor did they align with any other factors on the multilevel solution. As the last factor extracted on the traditional solution, it indicates that traditional solution provided an 'overextraction' of the *Group Learning* dimension (i.e., extracting more factors than what truly exists) resulting in a spurious dimension (Wood, Tataryn, & Gorsuch, 1999).

The items that form these factors. We next examined the items that saliently loaded on each factor across the solutions. Only one factor (*Effectiveness Motivation*) included the exact same five items on the multilevel and traditional factor solution. This

factor included children's internal motivations with items like "Tries activity when not forthcoming" and "Practices activities without prompting". Because both methods produced the same factors we could interpret the configural makeup of this factor as invariant across methods. This could have arisen since the items that comprised this factor in the traditional solution exhibited relatively weaker cross-loadings (below .20). This indicates that other factors on the traditional solution did not explain large proportions of the item variance (see Table 3). The distinctive nature of this factor could have made it less vulnerable to any model misspecifications (i.e., failing to model the teacher assessor) under the traditional factor analytic approach.

Three additional factors (i.e., *Acceptance of Novelty and Risk*, *Interpersonal Responsiveness in Learning*, and *Sustained Focus in Learning*) retained a similar, but not identical, factor pattern. If these factors differed with respect to their highest loading items (i.e., items with loadings closer to one), it could influence our conceptual interpretation of the factor since the higher loading items are better representations of the factor. This was not the case; only the lowest loading item(s) differed across the traditional and multilevel methods for these three factors, indicating that the factors have similar conceptual interpretations. *Acceptance of Novelty and Risk* contained four of the same items on the multilevel and traditional solutions. An additional item ("Previous attempts unsuccessful, still tries") loaded on the traditional factor for *Acceptance of Novelty and Risk*. This item evidenced the second weakest factor loading and therefore did not influence the general interpretation of the factor. *Interpersonal Responsiveness in Learning* contained the same nine out of ten items on both solutions. The multilevel solution uniquely had "Learns by accepting constructive feedback", and the traditional

solution uniquely included “Responds positively to assistance”. These two items represented the weakest factor pattern loadings and did not influence the general interpretation of the factors. Similarly, *Sustained Focus in Learning* had four of the same five items on both the multilevel and traditional solutions. The fifth item on the traditional solution was “Self-selects activity without direction” and, “Maintains essential role in small group” on the multilevel solution. Both items were the weakest factor pattern loadings on the solutions and similarly did not influence the conceptual interpretation of the factors.

The two remaining factors, *Strategic Planning* and *Demonstrated Engagement in Learning* had a larger departure in their item compositions between the traditional and multilevel solutions, suggesting potentially different conceptual interpretations. The multilevel *Strategic Planning* factor contained only three items whereas the traditional factor solution *Strategic Planning* had eight items. The three items that comprised the multilevel factor solution also had the highest loadings on the traditional factor solution. The three items on the multilevel factor solution included “Develops plan after considering the consequences”, “Basic understanding of cause and effect” and “Changes strategies when solution not working”. The traditional factor included these three items with the addition of five other items (e.g., “Self-corrects errors”, “Develops plan for multistep activity”). Since the highest loading items were unchanged across the solutions, it did not affect our conceptual interpretation of the *Strategic Planning* factors. However, because the traditional *Strategic Planning* factor contained a broader array of items, we could interpret it as providing more coverage of the *Strategic Planning* factor. For instance, it included being able to verbally demonstrate strategic planning skills

specifically with items ("Communicates that problems may have more than one solution" and "Verbalizes possible consequences"). Items for both *Strategic Planning* factors are included in Table 7 below.

TABLE 7

Items Comprising the Multilevel and Traditional Dimensions of Strategic Planning

Traditional	Multilevel
Develops plan after considering	Basic understanding of cause and effect
Basic understanding of cause and effect	Develops plan after considering consequences
Changes strategies when solution not	Compares new task to previous re: what
Self-corrects errors	
Compares new task to previous re: what	
Develops plan for multi-step activity	
Communicates problems may have more	
Verbalizes possible consequences	

The traditional and multilevel *Demonstrated Engagement in Learning* factors also evidenced larger configural differences than the previously compared factors. *Demonstrated Engagement in Learning* comprised ten items on the multilevel factor solution and three items on the traditional factor solution (see Table 8). Both solutions had the same three items of "Voluntarily demonstrates academic skills", "Verbalizes frustration and asks for help", and "Demonstrates pride in work products". Two of these items predominantly feature children's ability to demonstrate skills and pride in work products. We interpreted these items as preschool classroom teacher assessors responding to students' demonstrations of engagement in classroom activities. Since the highest loading items were unchanged across the methods, it did not affect our general conceptual interpretation of the *Demonstrated Engagement in Learning* factor. However, the difference in the number of items meant the multilevel *Demonstrated Engagement in*

Learning factor does capture a broader array of skills. Specifically, the multilevel factor additionally encompassed children's ability to verbally initiate requests of the teacher ("Asks teacher for a task", "Asks questions and shares ideas"), and verbally demonstrate critical thinking skills ("Communicates that problems have more than one solution", "Verbalizes possible consequences"). Items comprising both *Demonstrated Engagement in Learning* factors are included in Table 8 below.

TABLE 8

Items Comprising the Multilevel and Traditional Dimensions of Demonstrated Engagement in Learning

Traditional	Multilevel
Voluntarily demonstrates academic skills	Demonstrates pride in work products
Verbalizes frustration and asks for help	Verbalizes frustration and asks for help
Demonstrates pride in work products	Guesses even when unsure
	Verbalizes frustration but continues working
	Voluntarily demonstrates academic skills
	Verbalizes possible consequences
	Seeks answers by engaging with materials and people
	Asks questions and shares ideas
	Asks teacher for a task
	Communicates problems may have more than one solution

The larger configural changes in these solutions for both *Strategic Planning* and *Demonstrated Engagement in Learning* may be a result of the overextraction of the *Group Learning* factor that occurred in the traditional solution. Six of the new items on the multilevel version of *Demonstrated Engagement in Learning* (e.g., "Communicates that problems have more than one solution", "Verbalizes possible consequences") evidenced strong cross loadings with the *Group Learning* factor on the traditional solution. As such, these configural changes are likely due to the presence or absence of

the *Group Learning* dimension. Similarly, items that previously were salient loaders on the traditional dimensions for *Strategic Planning* (i.e., Develops plan for multi-step activity) also evidenced significant cross loadings with the traditional *Group Learning* dimension. Since both sets of items on the traditional *Strategic Planning* factor and traditional *Demonstrated Engagement in Learning* factors were highly correlated with the *Group Learning* factor, it made them more susceptible to changes in the final factor solution as a result of the absence of *Group Learning*.

The strength of the association between the items and factors. Our final task was to look at the strength of association between the items and the factors for factors with identical item compositions. Since only *Effectiveness Motivation* had the exact same items on the traditional and multilevel factors, we could only compare the relative strength of loadings for this factor. We found a similar magnitude of factor loadings on both solutions (average traditional loading = .54, average multilevel loading = .60). "Tries activity when not forthcoming" represented the highest loading item on both solutions. It loaded .64 on the multilevel solution and .70 on the traditional solution. The weakest loading traditional item, "Practices activities without prompting", loaded at .46 on the traditional solution and at .47 on the multilevel solution. Since the item loadings were so similar across solutions, we did not interpret this as a substantial change across the methods.

Analytic Results for Research Hypothesis 3: The Factor Structure of the LTLS Resulting from Multilevel Factor Analysis will Result in Significant Differences in the External Validity of LTLS Factors Compared to Those Resulting from

Traditional Factor Analysis, Evidencing one or more Factors with External Validity to Cognitive School Readiness Domains

Concurrent and predictive validity of all dimensions. First, we predicted academic ability measured concurrently in the school year. The seven traditional dimensions in a single regression model explained 27% of the variance in Mathematics, 17% of the variance in Listening Comprehension, 19% of Alphabet Knowledge and 19% of Vocabulary (i.e., Concurrent Validity variance explained by Model A; see Table 9). For academic outcomes recorded six months later, these seven traditional dimensions still explained 22% of the variance in Mathematics, 14% of the variance in Listening Comprehension, 15% of the variance Alphabet Knowledge and 16% of the variance in the Vocabulary outcome (i.e., Predictive Validity variance explained by Model A; see Table 9).

TABLE 9

Combined Variance Explained in Academics by Traditional and Multilevel Dimensions

Concurrent Validity	Mathematics	Vocabulary	Alphabet Knowledge	Listening Comprehension
Traditional Dimensions (Model A)	.27	.19	.19	.17
Traditional and Multilevel Dimensions (Model AB)	.27	.19	.20	.17

Predictive Validity	Mathematics	Vocabulary	Alphabet Knowledge	Listening Comprehension
Traditional Dimensions (Model A)	.22	.16	.15	.14
Traditional and Multilevel Dimensions (Model AB)	.23	.18	.15	.14

Note. All models were run in two stages. In the first set of models, all seven traditional dimensions predicted academic outcomes (i.e., "Model A"). This was followed by a second stage with the addition of the six multilevel dimensions added to the model (i.e., "Model AB"). This allowed us to assess how much additional variance the multilevel dimensions (Model AB) could explain above and beyond the traditional dimensions (Model A).

Adding all six multilevel dimensions to this model did not uniquely add to the variance explained from the combined predictive model (see Table 9 above Model AB results). The multilevel dimensions did not improve predictive capacity for three of the concurrent outcomes, namely Mathematics, Alphabet Knowledge or Listening Comprehension. However, these dimensions did explain an additional 1% of the variance in concurrent Alphabet Knowledge but Cohen's f^2 statistics for this increase was below the .02 threshold for small effects. In the predictive models for future academic performance, the multilevel dimensions explained an additional 1% and 2% of the variance in Mathematics and Vocabulary, respectively. Although, these values did not meet the f^2 threshold (.02) for small effects. The multilevel dimensions improved prediction by up to 2% of the model R^2, however, the improvement in R^2 did not meet small effect size benchmarks in relative variance explained. Collectively, this indicates that the multilevel dimensions are not likely to generate different practical inferences relative to external analyses using all of the dimensions from the traditional analysis.

Concurrent and predictive validity of individual dimensions. Next, we looked at the explanatory capacity of each of the traditional dimensions on their own. For both concurrent and predictive validity analyses, *Strategic Planning*, *Sustained Focus in Learning* and *Effectiveness Motivation* evidenced the strongest prediction for the four academic outcomes. They explained about 20% of the variance in concurrent Mathematics and approximately 12% of the variance in the other concurrent academic tests. *Demonstrated Engagement in Learning, Interpersonal Responsiveness in Learning,* and *Acceptance of Novelty and Risk* provided relatively weaker prediction of academic

skills. These dimensions accounted for approximately 6% to 16% of the variance in the academic outcomes (see Table 10).

TABLE 10

Concurrent and Predictive Validity for Variance Explained in Academics by Traditional Dimension

Concurrent Validity	Mathematics	Vocabulary	Alphabet Knowledge	Listening Comprehension
Effectiveness Motivation	.19	.10	.13	.10
Sustained Focus in Learning	.19	.11	.14	.12
Acceptance of Novelty and Risk	.14	.07	.09	.08
Demonstrated Engagement in Learning	.15	.11	.10	.11
Interpersonal Responsiveness in Learning	.16	.09	.12	.11
Strategic Planning	.26	.17	.18	.15

Predictive Validity	Mathematics	Vocabulary	Alphabet Knowledge	Listening Comprehension
Effectiveness Motivation	.16	.09	.11	.08
Sustained Focus in Learning	.16	.11	.10	.10
Acceptance of Novelty and Risk	.12	.06	.07	.06
Demonstrated Engagement in Learning	.12	.09	.08	.07
Interpersonal Responsiveness in Learning	.13	.10	.08	.09
Strategic Planning	.21	.15	.14	.13

Note . All calculations are variance explained by the traditional dimensions (Model A).

When adding the multilevel version of these dimensions to the model, improvements were observed for both *Demonstrated Engagement in Learning* and *Sustained Focus in Learning*. The biggest R^2 improvement was for *Demonstrated Engagement in Learning.* The multilevel factor of *Demonstrated Engagement in Learning* improved prediction for all four academic outcomes. It improved prediction by 5% in Alphabet Knowledge and Listening Comprehension. This analysis demonstrated stronger improvements in Vocabulary and Mathematics where the multilevel dimensions augmented prediction by 7% to 8% of the variance explained. Effect sizes on Cohen's f^2 ranged from .05 for Alphabet Knowledge to .09 for Mathematics. These f^2 values would be considered small effects (Cohen, 1988). The multilevel *Sustained Focus in Learning* model supplemented prediction from the traditional dimension of *Sustained Focus in Learning*. It explained an additional 3% of variance in Mathematics, Vocabulary and Alphabet Knowledge. These effects would also be considered small (f^2 = .02 to .03). Other multilevel factors improved prediction by approximately less than 1% and their effect sizes were below f^2 = .02.

The overall pattern for predictive validity mirrored findings for the concurrent validity analyses. When adding the multilevel factors, the new factor of *Demonstrated Engagement in Learning* improved prediction by roughly a change in R^2 of 4% in Alphabet Knowledge and Listening Comprehension, 6% in Vocabulary and 7% in Mathematics. An effect size was calculated at Cohen's f^2 = .07 for Vocabulary to Cohen's f^2 = .05 for Alphabet Knowledge. Similarly, *Sustained Focus in Learning* improved prediction in future academics by roughly 3%. These effect sizes are also considered small (see Table 11). Other multilevel factors improved prediction by

approximately less than 1% but these effect sizes do not meet the threshold convention for small improvements in model R^2 (Full Cohen's f^2 values are listed by predictor in Table 11).

The multilevel dimensions of *Demonstrated Engagement in Learning* and *Sustained Focus in Learning* explained more variance in external outcomes relative to the traditional dimensions. As such, these dimensions provide greater statistical power to detect small effects in future research with other academic outcomes or smaller sample sizes. Where these multilevel dimensions demonstrate conceptually relevant improvements in external validity for concurrent and future outcomes (such as *Demonstrated Engagement in Learning* providing increased capacity for predicting Vocabulary scores), these dimensions provide differential capacity to improve prediction to relevant external criterion outcomes.

TABLE 11

Cohen's $f2$ Improvement in Predictive Validity for Variance Explained in Academics by Dimension

Concurrent Validity	Mathematics	Vocabulary	Alphabet Knowledge	Listening Comprehension
Effectiveness Motivation	.00	.00	.00	.00
Sustained Focus in Learning	**.03**	**.03**	**.02**	.01
Acceptance of Novelty and Risk	.00	.00	.01	.00
Demonstrated Engagement in Learning	**.09**	**.07**	**.05**	**.05**
Interpersonal Responsiveness in Learning	.01	.00	.01	.01
Strategic Planning	.00	.01	.00	.00

Predictive Validity	Mathematics	Vocabulary	Alphabet Knowledge	Listening Comprehension
Effectiveness Motivation	.00	.00	.00	.00
Sustained Focus in Learning	**.03**	**.02**	**.02**	.01
Acceptance of Novelty and Risk	.00	.00	.00	.00
Demonstrated Engagement in Learning	**.06**	**.07**	**.05**	**.05**
Interpersonal Responsiveness in Learning	.01	.00	.00	.00
Strategic Planning	.00	.00	.00	.00

Note. Cohen's $f2$ effect size benchmarks are listed by predictor. Small effects (> .02) are bolded and shaded. These values represent the scaled additional variance explained in Model AB.

CHAPTER 4: DISCUSSION

Research findings strongly support that our most vulnerable young children living in poverty are more likely to be ready for school if they have a high quality early childhood education experience where educators are using the most scientifically based, child-centered assessments and curricula (Barnett, Weisenfeld, Brown, Squires, & Horowitz, 2016). Scientifically based preschool assessments are critical to guide and evaluate the extent to which comprehensive early childhood education interventions contributes to children's school readiness. The present research study draws attention to applying the most scientifically based, state-of-the-art methods from psychometric science to preschool assessment to ensure that our existing teacher-report measures of preschool domains of school readiness are of the highest quality.

Recent psychometric research has called for the use of complex, multilevel methods that solely rely on the variance of the child to develop and validate multidimensional teacher-report scales (Stapleton et al., 2016). These methods are designed to identify and account for unwanted assessor variance that if high enough may distort the construct validity of teacher-report scales. Traditional factor analytic approaches to empirically determine latent structures of important domains of school readiness do not account for assessor variance; however multilevel factor analytic approaches have been developed and used to increase precision when assessor variance is excessive (Muthen, 1994). These methods can identify where assessor variance is too high and control for it to yield more precise and child-centered dimensions of critical school readiness domains of functioning for use with our must vulnerable populations of children.

Heretofore, researchers have not yet applied these multilevel factor analytic methods to test existing multidimensional, teacher-report assessments in preschool early childhood research literature. Rigorous tests have not yet been conducted to determine if these multilevel factor analytic methods produce more valid dimensions of domains of school readiness compared to traditional factor analysis methods for preschool children living in poverty. The purpose of the present study was to conduct the first test of these methods on an important domain of school readiness-Approaches to Learning- with a valid multidimensional, teacher-report measure established using traditional factor analytic methods - the *Learning-to-Learn Scales* (LTLS). This test involved three sequential hypotheses designed to determine if multilevel factor methods make a substantial contribution over and above traditional methods in improving the validity of the LTLS for use with preschool, Head Start children. The following section will discuss the findings from testing each of these hypotheses.

Discussion of Hypothesis 1: The LTLS, as a Teacher-Report Measure of Approaches to Learning, has Items that Contain a Significant Amount of Assessor Variance Warranting the use of Multilevel Factor Analytic Methods

The first hypothesis tested the proposition that a significant amount of the item variance in the LTLS is associated with the classroom teacher assessing the students (i.e., assessor variance) rather than the students themselves. Specifically, it was hypothesized that greater than 5% assessor variance would be found among the items on the LTLS. The analyses indicated an average of 21% assessor variance among the items on the LTLS, which is over four times the amount of assessor variance that the research

literature requires to signal the need to use multilevel statistical methods (i.e., 5%, Pornprasertmanit et al., 2014).

Studies of preschool and elementary school teacher-report assessments consistently report that the amount of assessor variance ranges from 22% to 69%, with an average of approximately 33% of the variance attributable to the teacher rater (Barghaus et al., 2017; Howard et al., 2017; Kim et al., 2016; Waterman et al., 2012). These findings indicate that teacher-report of young children typically generates high levels of assessor variance that greatly exceed the 5% threshold to be concerned (Waterman et al., 2012). Teachers often must complete assessments in addition to numerous other responsibilities that leave them with little time and resources to focus on student differences when filling out assessment forms (NRC, 2008). This results in assessment data with less precisely defined differences between children assessed by the same teacher, thereby resulting in a large amount variance associated with the teacher assessor (Waterman et al., 2012). These findings suggest that the LTLS, like other early childhood teacher-report assessments violate the assumption made by traditional analytic methods that there is negligible assessor variance present in the child assessment data.

The significant amount of assessor variance (21%) identified in the LTLS is over four times the 5% threshold suggesting use of multilevel analysis; however, it is lower than that of other teacher-report assessments used in preschool and elementary settings (Waterman et al., 2012). In prior studies, higher amounts of assessor variance were identified in assessments used as part of routine practice and contained in administrative records, while lower amounts were found as part of university-led research and validation studies of new assessments (e.g., Goldstein & McCoach, 2012). The LTLS was

developed as part of a university research project that required careful and rigorous data collection practices. In this study, teachers knew they were rating students on the LTLS as part of a research study and were provided incentives for completing the assessments (see McDermott et al., 2011). Because of this, teachers may have put more time and effort into considering individual differences when completing the LTLS for each child in their class (McDermott et al., 2011). The difference in time, resources, and emphasis on attention to detail that teachers face when completing assessments for routine administrative practice versus a research study may help explain why the assessor variance in the current study of the LTLS was lower than what is typically observed in early childhood teacher-report assessment.

The findings from testing the first hypothesis reveal excessive assessor variance in the teacher-reported LTLS assessment. These findings bring into question the scientific integrity of using traditional statistical methods to validate the latent structure of the LTLS. As indicated in the psychometric literature this level of assessor variance calls for the application of multilevel factor analysis methods to account for the assessor variance found to provide a more precise *child-centered* assessment of children's Approach to Learning abilities observed in the preschool classroom.

Discussion of Hypothesis 2: The use of the more Empirically Defensible Multilevel Factor Analytic Method will Produce a Distinctively Different Latent Factor Structure of the LTLS than one Produced by a Traditional Factor Analytic Method

Based on previous studies of multilevel factor analysis (Kim et al., 2016), it was hypothesized that the multilevel method would result in a different latent factor structure of the LTLS compared to the traditional method. Results showed that not only was a

difference found, but that the difference was the most severe type of difference that can emerge when making this comparison -- a change in *both* the number of dimensions identified and the nature of the dimensions. The number of LTLS dimensions dropped from seven dimensions derived from the traditional analysis to six dimensions resulting from the multilevel analysis. In addition, the multilevel factor solution produced a qualitatively different *Demonstrated Engagement in Learning* dimension.

The multilevel analysis did not include the *Group Learning* dimension that was identified in the traditional approach. There are two plausible explanations that *Group Learning* was not identified by the Multilevel Factor Analysis. First, the *Group Learning* factor could have been excluded because it was the last factor extracted in the Traditional Factor Analysis. Factor analysis as a statistical methodology produces latent factors that are always generated in order of how much variation they explain in the items; the factors that are extracted first are the "strongest"; and the factors extracted later in the analysis are "weakest" because they explain smaller amount of variance (Cattell, 1966). Weaker factors, like *Group Learning*, could more susceptible to changes when multilevel factor analytic procedures are applied since they only explain minor portions of item variance, however, this should be investigated in future research. Some research indicates that methodological choices, like which factor rotation to use, may lead to dropping or keeping the weakest factor (Finch, 2011). It is plausible that the decision to retain the "weakest" *Group Learning* factor would be affected by the application of Multilevel Factor Analysis but this warrants further investigation in an emerging field of research on Multilevel Factor Analysis applied to teacher-report child assessment.

Second, *Group Learning* may have emerged because the items on the *Group Learning* factor were in close proximity to one another on the LTLS administration form (see Appendix D). As an analysis of covariance, factor analytic methods can produce spurious factors when items are artificially highly correlated because they were included next to each other on an assessment form (Reise, Waller & Comrey, 2000). For example, once a teacher rates a student low on an item assessing a particular construct, they are more likely to continue to rate the student similarly low on other items assessing that construct if they are clustered together on the assessment form (Hurd, McFadden, Chand, Gan, Menill & Roberts, 1998). This is why it is generally recommended to intersperse items that assess a construct (Shrieshein & DeNisi, 1980). Once the multilevel factor analysis statistically accounted for assessor variance, the *Group Learning* items no longer shared enough variance to load onto a factor as they did in the traditional method. This finding that the *Group Learning* dimension may be a spurious factor points to the need to revise the LTLS administration form so that these items are no longer grouped together. Future work could then test whether the *Group Learning* dimension emerges once items are interspersed across the administration form.

The multilevel *Demonstrated Engagement in Learning* dimension was found to be more substantial than the traditional dimension with more items that more robustly define the nature of children's behaviors that demonstrate engagement in learning. The traditional dimension encompassed just three items: "Demonstrates pride in work products", "Voluntarily demonstrates academic skills", "Verbalizes frustration and asks for help." Two of these items represent children's initiative and curiosity in the classroom and one item expressed pride in what their initiation and agency produced in

the classroom. These three together demonstrate both children's active engagement as captured by Head Start's Approaches to Learning subdomain of Initiative and Curiosity (Goal P-ATL 10-11; ACF, 2015a) and confidence in one's own skills and positive feelings about self as reflected in the Social and Emotional Development subdomain of Sense of Identify and Belonging (Goal P-SE 10).

The multilevel *Demonstrated Engagement in Learning* dimension included seven additional items to the original three which highlight verbal demonstrations of initiative and curiosity in classroom activities ("Seeks answers by engaging with materials and people", "Asks questions and shares ideas", "Asks teacher for a task", "Guesses even when unsure", "Verbalizes frustration and continues working", "Verbalizes possible consequences", "Communicates that problems may have more than one solution"). These items in combination with the previous three represent important ways that children demonstrate how they are engaged in productive independent activities, communicate choices to adults, willingly participate in challenging activities, and express pride in what their initiative and curiosity produces. The multilevel factor analytic approach resulted in a more robust dimension of engagement than the traditional factor analysis method with more aspects of children's *Initiative and Curiosity* assessed consistently by their teachers.

For young children, *Initiative and Curiosity* has been operationalized as an openness toward new challenges and the "impulse towards better cognition" (Kagan, Moore, & Bredekamp, 1995). Children are naturally interested in learning more about how the world works, and they drive their own development by taking initiative to seek out new experiences becoming 'active agents' for their own learning (Bronfenbrenner & Morris, 1998). By now having a more robust dimension of children's initiative and

curiosity allows teachers to more comprehensively relay their students' ability to seek out tasks that push their competencies and allow them to mature (Kagan, Moore, & Bredekamp, 1995). It is not surprising that kindergarten teachers believe that curiosity is a more important predictor of school readiness than knowledge competencies such as counting or understanding of the alphabet (Jirout & Klahr, 2012). Indeed, children's initiative and curiosity has been found to account for nearly a third of the variance in preschooler's performance on academic assessments (Jirout & Klahr, 2012; Dobbs, Doctoroff, Fisher, & Arnold, 2006).

The inclusion of verbal demonstrations of initiative and curiosity in the multilevel *Demonstrated Engagement in Learning* dimension captures the verbal communications by which teachers and children most commonly interact. Children's ability to vocalize their needs and ask questions of teachers in the classroom helps them engage in scaffolded interactions with teachers (Halle & Darling-Churchill, 2016) and make gains in language and academic skills (Dickinson & Porche, 2011). Thus, the multilevel *Demonstrated Engagement in Learning* dimension is better able to monitor children's readiness for school because it more fully captures verbal demonstrations of initiative and curiosity in the preschool classroom with the addition of these new seven items.

The results of this first test of multilevel factor analysis evidence a major difference when accounting for the assessor variance: a change in the number of factors with qualitative distinctions between the factors. This is a more severe difference than generally what has been found in studies comparing multilevel and traditional factor analytic approaches (Kim et al., 2016). While very few studies have applied multilevel factor analysis to teacher-report child assessment, two documented a difference in the

number of factors extracted and conceptual reinterpretations of the factors that emerged when comparing a traditional and multilevel factor analytic approach (Peters, Algina, Smith & Daunic, 2012; Barghaus, LeBoeuf, Fantuzzo, Brumley, & Coe, 2017).

Peters et al., (2012) studied the use of multilevel factor analysis to teacher-report of elementary students' Executive Functioning. Two factors were hypothesized to be captured by the instrument under a traditional analytic framework but three dimensions emerged at the child-level of analysis. This unanticipated factor was called "Emotion Regulation Index" that was previously hypothesized to be part of their "Behavioral & Emotional Self-Regulation" factor. Their findings indicated greater differentiation of children's skills when using multilevel factor analysis as opposed to traditional factor analysis. Similarly, Barghaus, LeBoeuf, Fantuzzo, Brumley, & Coe (2017a) found that one less factor emerged instead of the hypothesized three factors when they applied multilevel factor analysis to a teacher report assessment of kindergarten children's engagement behaviors. The hypothesized general factor of engagement, previously found in a traditional factor analysis, did not emerge in a multilevel analysis. However, two of the factors, *Academic Engagement* and *Social Engagement*, remained. Like these two studies, a multilevel factor analysis with the LTLS data produced a different number of factors and the items that comprised those factors, which in turn influenced our conceptual understanding of the factors. Such profound differences between traditional and multilevel methods underscores how important it is to consider multilevel factor analysis when developing and validating teacher-report child assessments.

These dimensionality findings demonstrate the necessity of multilevel factor methods for the development and refinement of preschool teacher-report measures like

the LTLS when significant amounts of assessor variance are found. The significant differences found in the number and nature of the LTLS factors support finding significant differences in external validity favoring the multilevel factor analysis method. The findings suggest multilevel factor methods will result in differences in the validation of teacher-report child assessments in early childhood education research. The next test of this concerning possibility is presented in the Hypothesis 3 section below.

Discussion of Hypothesis 3: **The Factor Structure of the LTLS Resulting from Multilevel Factor Analysis will Result in Significant Differences in the External Validity of LTLS Factors Compared to Those Resulting from Traditional Factor Analysis, Evidencing one or more Factors with External Validity to Cognitive School Readiness Domains**

The final hypothesis tested whether the differences in the multilevel factor structure of the LTLS would result in differences in the strength and pattern of external validity evidence for children's academic outcomes. In particular, it was hypothesized that the multilevel factors would explain more variance in children's outcomes than the traditional factors. Findings revealed improvements in the external validity of the multilevel dimensions over the traditional dimensions with the most striking difference in for the *Demonstrated Engagement in Learning* factor.

The multilevel *Demonstrated Engagement in Learning* factor improved in the prediction of academic outcomes by roughly 45% to 80% over the traditional dimension. The improvement in external validity was most evident for children's vocabulary skills five months later where the multilevel dimension predicted 80% more variance than the traditional dimension. This is not surprising given that the multilevel *Demonstrated*

Engagement in Learning factor included more items representing a broader array of skills, including children's vocal engagement in the classroom, compared to the corresponding traditional factor. Prior studies have identified that preschoolers who are more vocally engaged in the classroom have a larger vocabulary in elementary school (McClelland, Morrison & Holmes, 2000; McClelland et al., 2007). Vocal engagement in the classroom likely improves children's vocabulary over time because young children who talk more tend to elicit more language input from teachers (Whorrall & Cabell, 2016). Greater exposure to conversations with teachers provide more opportunities to learn new words and additional meanings of known words (Cabell, Justice, McGinty, DeCoster, & Forston, 2015). Thus, the current findings suggest that the addition of verbal engagement skills—captured by the multilevel *Demonstrated Engagement in Learning* dimension—within the classroom context may help children's prospective vocabulary development.

The current study is the first to compare the external validity of a multilevel approach to a traditional approach in a teacher-report assessment of preschool children. However, we can situate the statistical magnitude of the improvement from an application of multilevel methods with the LTLS against two prior studies using multilevel factor analysis to improve teacher-report assessment of children in Kindergarten (Howard et al., 2016; Barghaus et al., 2017). Howard and colleagues (2016) and Barghaus and colleagues (2017) found that the correlations between multilevel factors and later academic outcomes were 25% to 200% stronger than correlations between the traditional factors and academic outcomes. In particular, Howard et al. (2016) found that multilevel methods provided a small 25% improvement in the strength

of the association between school readiness in Kindergarten and later academic outcomes compared to correlations based on traditional statistical methods (Cohen, 1992). Moreover, Barghaus and colleagues (2017) found that multilevel factors of Kindergarten classroom engagement explained an average of three times more variance in later academic outcomes compared to the traditional factors. Our finding of approximately a 45% to 80% improvement for predicting preschool children's outcomes fell in between the effects observed by Barghaus and colleagues (2017) and Howard and colleagues (2016) for children in Kindergarten.

In sum, our findings suggest that multilevel methods provide increased capacity to explain variance in important external outcomes. The greatest improvement in predictive ability was seen for the *Demonstrated Engagement in Learning* factor, which was likely due in part to the fact that the multilevel version of that dimension was more robust and comprehensive than the corresponding traditional factor. The observed improvements from application of multilevel methods to the LTLS—a teacher-report assessment in preschool--were generally consistent with what have been previously reported in studies of teacher-report assessments in Kindergarten.

Limitations and Future Research

This study provided the first empirical test of the impact of accounting for assessor variance in determining the validity of a teacher-report child assessment for preschool students. By basing dimensionality solely on the child variance, multilevel statistical methods provided more precise dimensions of the LTLS and stronger external validity than were found with the traditional methods. While this was the first empirical test of comparing multilevel factor analysis methods with traditional methods in the

assessment of preschool children, there are limitations of this single study that point to the need for additional research. First and most obvious, these findings are only a single test of the impacts of assessor variance in preschool assessment and therefore more studies are needed. Second, extending beyond external criteria measures from the *Cognition* and *Language and Literacy* domains captured by the Learning Express assessment may reveal more benefits of using a multilevel model to uncover external validity evidence for other measures of *Approaches to Learning* to *Social and Emotional Development*. Third, future research comparing multilevel and traditional factor analysis methods should utilize multiple observers of children's Approaches to Learning behaviors to provide a more precise control of assessor variance.

Only one study.

The findings of this study represent only a single test of multilevel factor analysis and these findings are conditional on the employed sample (Thompson, 2002). Campbell and Stanley (1963) caution against considering even a well-designed single study as definitive evidence of a broader phenomenon, suggesting instead that single studies be viewed as a "path towards accumulating knowledge". To ensure a rigorous test of multilevel factor analysis, the same sample was used for both analyses and held many of the design features the same (e.g., the statistical estimation method and the factor rotation procedure). This strengthens our inferences in comparing the multilevel dimensions to the traditional dimensions. However, it is unknown if and to what degree the differences observed between the multilevel and traditional method are specific to the sample of Head Start students in a large, high-needs school district with financial limitations for

professional development. Replication is therefore important to confirm the general conclusion of these study findings (Makel & Plucker, 2014).

To ensure the generalizability of these findings, multilevel factor analysis should be used with other samples of teacher and student participants. Past research has revealed the prevalence of assessor variance in widely implemented early childhood assessments (e.g., Waterman et al., 2012), but more work is needed to look at the variation within assessments across samples or across assessment contexts for widely-used assessments. For instance, assessor variance estimates could change based on the response context (i.e., controlled research studies versus routine administrative assessments). The multilevel factor method should be applied for these assessments under these different conditions to test for confirmation of a broader phenomenon. Such work would require no additional burden on teacher assessors; instead, multilevel factor analysis is a way to strengthen early childhood assessment at the point of statistical analysis, which makes it an appealing and feasible direction for future work.

Future research should employ multilevel methods to evaluate assessments of other major domains of child functioning (e.g., *Cognition*) for preschool children, which may reveal a different pattern of findings. There is some preliminarily evidence that asking teachers to rate children's Approaches to Learning may result in greater amounts of assessor variance than when they rate other, more concrete, domains of child functioning (Howard et al., 2017). Thus, future studies of other domains of child functioning may find less severe departures between traditional and multilevel factor analyses than what was observed here because Approaches to Learning may be particularly prone to assessor variance (Waterman et al., 2012). Such future research

would help identify the domains of child functioning for which assessor variance is likely to be a concern and the use of multilevel methods should therefore be considered.

Only cognition and language and literacy domains were used.

Psychometric validation typically consists of comprehensive convergent validity evidence sampling from major domains of child functioning (Campbell & Fiske, 1959; Bulotsky-Shearer & Fantuzzo, 2004). However, the current study only used measures from the *Cognition* and *Language and Literacy* domains, as implemented in the Learning Express child assessment, for external validity criterion by which to compare the traditional and multilevel dimensions of the LTLS. The present research could be further extended by utilizing additional external validity criteria beyond the Learning Express. The Learning Express is referenced to state and federal learning frameworks and therefore represents an important set of outcomes relevant to Head Start students (McDermott et al., 2011). Although the Learning Express measures outcomes that are theoretically and empirically related to Approaches to Learning, it itself is not a measure of Approaches to Learning like the LTLS. No measure of Approaches to Learning administered by independent assessors (i.e., not teachers) was available when the current study was conducted. Therefore, future studies should test whether differences emerge between the traditional and multilevel dimensions of the LTLS when predicting relevant Approaches to Learning skills (Hyson, 2008).

Future work could use other external observational assessments of children's Approaches to Learning skills as additional validation criteria. For example, the inCLASS is an observational measure of preschooler's Approaches to Learning that can be completed by a trained observer (Downer, Booren, Lima, Luckner, & Pianta, 2010).

The inCLASS measures three major domains: *Teacher Interactions, Peer Interactions,* and *Task Orientation.* The *Teacher Interactions* domain includes ratings of the quality of child-teacher interactions and the child's use of language to engage with the teacher, and the *Task Orientation* domain assesses the child's engagement with classroom tasks and activities. Observer ratings on the inCLASS *Teacher Interactions* and *Task Orientation* scales could be used as external measures of children's engagement with their classroom and provide external validity evidence for the multilevel *Demonstrated Engagement in Learning* dimension of the LTLS.

Additionally, future studies could employ direct child assessments of domains of Approaches to Learning like the Head-Toes-Knees-Shoulders Task (Ponitz, McClelland, Matthews & Morrison, 2009). The Head-Toes-Knees-Shoulders Task measures children's behavioral regulation including attentional focusing and inhibitory control. Measuring such important preschool classroom skills could be used to validate the multilevel *Sustained Focus in Learning* dimension of the LTLS which also purports to measure children's sustained attention and ability to inhibit distracting behaviors. The Head-Toes-Knees-Shoulders task could demonstrate statistically higher external relations with the *Sustained Focus in Learning* dimension and relatively lower with *Demonstrated Engagement in Learning* dimension which would provide convergent and divergent validity evidence for the LTLS. Using additional assessments of Approaches to Learning would provide a more direct comparison for validation of the LTLS and could illustrate more improvements in external validity of the multilevel dimensions over the traditional dimensions.

Multilevel analysis limits the use of assessor variance.

The multilevel analysis used in this study identified and removed all assessor variance from the LTLS items, but some of that assessor variance may be informative upon further analysis. Teacher-report instruments, like the LTLS, provide a child's assessment score based on a single teacher's perspective. This is problematic because it is unclear whether variation in children's scores is attributable to true individual differences or assessor bias (Waterman et al., 2012). For example, if one teacher rates her students on average higher on the LTLS than another teacher rates her students, then it is unclear whether the first teacher's students are in fact higher in school readiness, or if this reflects that one teacher tends to rate students more optimistically than the other. The multilevel analyses used in this study cannot distinguish between true differences in classrooms of students and teacher's own biases because all of the variance associated with the teacher assessor is removed.

Having multiple informants providing ratings on each child could provide a more precise way of isolating assessor variance than multilevel analyses of a single teacher's ratings (Jasyasinghe, Marsh, & Bond, 2003). A better practice than single teacher-report is to use multiple informants (e.g., multiple teachers, teacher's aides, or extramural assessors) because it better allows methodologists to distinguish between teacher bias and true classroom differences (NRC, 2008; Konold & Cornell, 2015). For example, if multiple observers indicate consistent classroom differences, then it is more likely to that those differences are real rather than attribute rater biases. Thus, having a classroom assistant or an extramural assessor provide a second rating of children's classroom

behavior would allow for more analyses that may provide insight into distinguishing rater bias and true classroom differences (Konold & Cornell, 2015).

In sum, the findings of the present study demonstrated that assessor variance is a threat to the validity of teacher-report assessments of multidimensional constructs of school readiness of preschool children that must be addressed. High levels of assessor variance were found on the LTLS, a multidimensional, teacher-report of Approaches to Learning skills, that affected the internal and external validity of the measure with a population of urban Head Start preschool children. When these high levels were controlled for using Multilevel Factor Analysis methods, a significant difference was found in the factor structures in the external validity of the LTLS. The next section will consider the appropriate short-term and long-term policy and practice implications of this research for preschool assessment in general and particularly in Head Start.

Implications

The results of this study underscore the importance of making visible how assessor variance can have an adverse impact on the validity of preschool teacher-report assessments. The current results also demonstrate how researchers can use state-of-the-art, multilevel psychometric methods to account for assessor variance and lessen its threat to the validity of these important measures. Given the importance of early school readiness intervention for young children from low-income households, it was important that this first scientific test of the impact of assessor variance be conducted on a multidimensional, teacher-report assessment, the LTLS, that was intentionally validated for use with Head Start children to support a critical domain of school readiness. The LTLS is currently the best documented multidimensional assessment for preschool

Approaches to Learning in the literature. Though this is only a single investigation, the results have important policy and practice implications. This section will describe first, how these findings can be used to make the early childhood education field aware of the threat of assessor variance to the validity of assessments being used by preschool teachers in Head Start; and second, what concrete steps can be taken, mindful of this threat, to improve preschool teacher-report assessments and their use by teachers and administrators.

National study of widely used preschool teacher-report assessments.

Historically, many researchers were unaware of the importance of multilevel methods until the pioneering work of Raudenbush and his colleagues (Raudenbush & Bryk, 1999). Their research showed that traditional methods do not account for classroom or school-level variance. Ignoring the variance associated with children's shared educational context (e.g., classrooms, schools) violates a key assumption made by traditional methods that observations are independent. This can bias estimates of standard errors, and thereby produce misleading conclusions about the statistical significance of effects. Raudenbush and Bryk (2002) went on to demonstrate that accounting for the classroom-level or school-level variance by using multilevel methods provides more precise estimates. Now, multilevel methods are standard research practice in situations where children are nested within teachers, classrooms or schools—such as the case in the current study. The present preschool study extends the few available studies examining assessor variance among teacher-report elementary school assessments by showing that the data collected on the LTLS far exceeded the 5% assessor variance threshold that research has shown can undermine the validity of the instrument unless multilevel

statistical techniques are used (Pornprasertmanit et al., 2014). To determine the extent of this problem in widely used teacher-report assessments in Head Start, more research is needed to evaluate and remedy the threat of assessor variance to the validity of multidimensional assessment used in Head Start by making full use of multilevel methods.

The first step towards making visible the extent of the problem that assessor variance poses in Head Start is to identify the teacher-report assessments that are currently widely used in Head Start. A model for doing this was established by the National Research Council [NRC] in their investigation early childhood assessments (NRC, 2008). Commissioned by Congress, the NRC sought to identify widely used teacher-report assessments of school readiness across multiple domains of school readiness competencies in Head Start. They first, searched multiple, independent online scholarly research databases (e.g., PSYCINFO, ERIC) and online databases that included additional instruments[8]. They also followed scientific guidelines to include "grey literature" including recent print and electronic reviews including compendia documents and technical summaries[9] to identify instruments that otherwise might not be recorded in the established online scholarly research databases (Cooper & Hedges, 1994; Siddaway, Wood & Hedges, 2018). Specifically, the NRC report included all assessments identified from these research database and "grey" sources in service of producing an inclusive

[8] Databases such as Buros Mental Measurements Yearbook, Buros Center for Testing Database, National Institute for Early Education Research Database, Educational Testing Service TestLink, and DPPeds have all been used in the NRC report. Recent study by the Penn Child Research Center also includes information on widely-used assessments and their existing validity evidence (Barghaus et al., 2017).

[9] Reports produced by The National Children's Study, The National Early Childhood Technical Assistance Center, Child Trends, The Center for Educational Measurement and Evaluation, and Mathematica Policy Institute have been used for the NRC report.

summary of available school readiness assessments "that have been widely used to reflect status or progress in that domain" (p. 87; NRC, 2008). The NRC report then organized the measures by the five domains of school readiness identified by the National Education Goals Panel and method of data gathering (i.e., direct assessment, questionnaire, observation, or interview; pp. 4, 120 - 144; NRC, 2008). This provided an overview of the number and types of assessments available to assess each school readiness domain. Finally, they provided a brief description of the databases and sources that provide additional information on psychometric characteristics of the instruments, but the NRC itself did not review the reliability and validity of the instruments (NRC, 2008). Although this review provides an excellent model for identifying a comprehensive collection of assessments used in early childhood that could be followed to identify specifically teacher-report measures, it did not report the reliability or validity of these instruments and did not apply the most rigorous psychometric science to evaluate validity for use.

Another effort commissioned by the Office of Planning Research and Evaluation entitled, *Understanding and Choosing Assessments and Developmental Screeners for Young Children Ages 3-5: Profiles of Selected Measures* attempted to specifically identify and report psychometric information for widely used measures in Head Start and early childhood (Halle, Zaslow, Wessel, Moodie, & Darling-Churchill, 2011). The purpose of this project was to assemble a compendium of measures to help Head Start and other early childhood education administrators review existing measurement tools and highlight areas in which the early childhood field is lacking information on reliability and validity of early childhood assessments and developmental screeners. They reviewed

information on the common indicators of reliability (inter-rater, test-retest, internal consistency), as well as *some* indications of validity (content validity, construct validity, convergent validity and predictive validity). However, no evidence of response process validity, validity evidence of the internal structure or consequential validity was reviewed. Halle and colleagues (2011) then independently summarized the quality of the reported evidence for reliability based on published psychometric guidelines, which are presented in Appendix B. This resulted in only 18 instruments including 5 teacher-report measures reviewed, however, it provided no direct evaluation of the quality of the measures. The report only compiled what other studies, principally those conducted by the instrument developers, had reported about the measures. As such this compendium fell short in that it was not comprehensive review of the most widely used measures across school readiness domains, like the NRC review, and it did not apply the most scientifically rigorous standards and psychometric methods to the measures reviewed.

Therefore, we need a comprehensive model like the one applied by the NRC that specifically targets widely used teacher-report measures across school readiness domains but one that also rigorously evaluates them apart from the claims of the developers and publishers. The Institute of Education Sciences (IES) established an excellent model of the Preschool Curriculum Evaluation Research (PCER) that could be applied to assess assessor variance and address its the threat to the validity of these assessments. PCER was a large-scale effort to investigate the scientific integrity of widely used curricula for preschool children (PCER, 2008). PCER addressed the lack of rigorous, systematic, randomized evaluations of preschool curricula by supporting small-scale efficacy evaluations using a common protocol and a standardized research randomized control

trial design (PCER, 2008). A peer-reviewed grant competition was created to have research teams around the country submit proposals to evaluate curricula "of [the evaluators] choosing" (p. xxxii; PCER, 2008). All of the proposals were assessed for key scientific standards including a standardized method of randomization, teacher training, implementation of the curricula, training of the assessors and collection of baseline and post-intervention and measures based on the latest scientific guidelines outlined by the IES request for proposals in 2002 (pg. xxxviii, NRC, 2008). Rather than one overall evaluation, PCER contains individual evaluations for each curriculum with common study designs to ensure replicability of the approach. Research teams collected data using a predetermined research protocol with planned fidelity of implementation measures that was used to justify the scientific integrity of each evaluation.

A similar approach could be used to systematically investigate widely-used teacher-report preschool assessments using the state-of-the-art multilevel methods demonstrated in the current study using their common research protocol that includes standardized research design and measures of fidelity of implementation. This would involve research teams who were not involved in the initial development or validation of an instrument using a standardized research protocol to test whether the amount of assessor variance present in a teacher-report assessment exceeds the five percent threshold to be concerned. Where significant assessor variance is found, the instruments could then be assessed using standardized multilevel analytic methods to investigate the impact of assessor variance using a standardized research protocol that also includes measures of fidelity of implementation to ensure reliable results. As done in the PCER study, the researchers should ensure fidelity of their implementation of the analytic

procedures such as by having their statistical code approved by a methodological team to ensure consistent statistical analyses.

Finally, once a complete set of analyses have been concluded for each teacher-report child assessment, dissemination of this information should occur through multiple channels. This should include: (1) leading early childhood journals (e.g., *Early Childhood Research Quarterly*; *Early Education and Development*), (2) presentations at academic and professional conferences dedicated to advancing scientific knowledge about education (e.g., American Education Research Association, Society for Research in Child Development), and (3) presentations directly to Head Start grantees (administrators and practitioners) by the Administration for Children and Families through its various national and regional dissemination channels. This important review would likely stimulate some important practical short-term and long-term responses to enhance the use of scientifically-based assessments in Head Start.

Application of the findings of the national study.

There are short-term responses that will be necessary to improve the use of teacher-report assessment for children in Head Start. Strategic short-term responses should prioritize reducing the presence of assessor variance in the existing widely-used early childhood teacher-report assessments. This could be targeted by new professional development initiatives that offer to support current Head Start administrators and teachers' understanding and use of existing teacher-report assessments. These efforts should be informed by principles from the field of Implementation Science which encourages researchers and stakeholders, like Head Start teachers, to collaboratively develop a plan to better implement evidence-based recommendations, such as using

child-centered assessment to inform instruction (Forman et al., 2013). There is research from Implementation Science showing that one-time training, a common professional development approach in Head Start, is not enough to ensure behavior change; rather, professionals require additional strategies such as ongoing consultation, incentives, and a supportive organizational culture to sustain their implementation of an evidence-based practice such as administering scientifically-based assessments (Stirman, Gitner, Langdon, & Graham, 2016).

In order to increase their acceptability and feasibility, teacher stakeholders should be involved in designing such professional development and implementation strategies to enhance early childhood teachers' understanding of measurement issues that arise in teacher-report assessments, such as assessor variance and its impacts on the measures' ability to provide valid information on children's school readiness. Ultimately, increasing understanding of these issues through implementing ongoing professional development would better foster a culture of teachers using assessment in a child-centered manner as an intrinsic part of the teaching and learning process – an important component of Head Start's own statement on effective educator practices (ACF, 2015a).

These efforts in assessment-focused professional development could be funded through Head Start's Technical Assistance and Training (TA/T) system as well as through initiatives put on by assessment publishers. Head Start's TA/T system dedicates up to three percent of total Head Start funding to improve "program quality" including the support of staff training and professional development (Kaplan & Mead, 2017). Previously, this mechanism has focused its efforts on helping Head Start staff attain bachelor's degrees. However, attaining a bachelor's degree is not enough to ensure that

Head Start teachers know how to best use assessment in a child-centered manner to minimize assessor variance and guide instruction. Head Start should allocate some TA/T funds to design and provide such specialized training and ongoing support in classroom assessment. In particular, Head Start could allocate their TA/T training funds from grant appointments to the National Center on Early Childhood Development, Teaching, and Learning to improve teacher professional development on the topic of assessment. These funds could support involvement of teacher stakeholders in the design and implementation of acceptable trainings, ongoing consultation, and incentives to promote best practices of child-centered assessment.

In addition to professional development implemented through Head Start's own funds, Head Start programs can request training from assessment publishers, for example, as part of the package that is purchased by existing Head Start programs. A benefit of this approach is that publisher-provided training would be specific to instruments that teachers are using in their practice. Such concrete, instrument-specific collaborative professional development would likely be acceptable to teachers and feasible to complete within a short amount of professional development time (Garet, Porter, Desimone, Biman & Yoon, 2001). This targeted professional development would help clarify the meaning of items and the broader constructs they are meant to assess through discussion and hands-on practice as these features have been identified in review of effective professional development practices (Garet et al., 2001; Guskey, 2003).

The impact of these national efforts to improve the use of improved widely used measures provides an excellent opportunity to investigate whether these training approaches are effective . Short-cycle evaluation studies could examine whether the

professional development initiatives had the intended effect of improving the validity of teachers' responses and reducing assessor variance in the items of widely-used teacher-report assessments. This could be accomplished at low cost as teachers' ratings on child assessments are available in Head Start administrative data; thus, no new data collection would be needed. Such program evaluation research could be supported through the Low-Cost, Short-Duration Evaluation of Education Interventions grant program initiated by the U.S. Department of Education Institute of Education Sciences. This program supports randomized controlled trials to evaluate the impact of education interventions conducted for $250,000 or less, completed within two years, and relying on administrative data for outcome monitoring. This funding mechanism could support the implementation and evaluation of many different professional development strategies (e.g., teachers randomized to training-only, training with ongoing consultation, or training with ongoing consultation and incentives) at low cost. Evaluating the effectiveness of such professional development strategies and modifying them as needed would support teachers and improve the quality of data collected on children.

Long-term policy recommendations.

Assessment scientists and publishers will need to focus on longer-term efforts to minimize the impact of assessor variance on teacher-report assessment in early childhood by increasing the validity of new assessments. This will require efforts at the point of developing new teacher-report assessments based on the latest Standards for Educational and Psychological Testing with specific attention to assessor variance in teacher-report child assessment (AERA et al., 2014). For instance, best psychometric practices outlined in the Standards for Educational and Psychological Testing encourages rigorous

assessment development methods to refine items prior to validation testing. Such a priori methods include conducting qualitative cognitive interviews to gather feedback from stakeholders to refine item wording and administration format/instructions (Dewalt et al., 2007). This would involve identifying factors that result in assessor variance, such as item wording that is interpreted differently among teacher assessors, or similar items that are placed near each other on the rating form (Downing, 2003). This would strengthen validity of the assessments by ensuring that teachers understand the constructs assessed by the items in the same way that the researchers intend.

After development of early childhood teacher-report assessments, psychometric validation research should require reports of assessor variance present in the items and, when there is greater than 5%, require the use of multilevel methodology in the first phase of validation. Such work can follow the example of multilevel analysis presented in Hypothesis 2 of the current study. The factor structures resulting from multilevel methods should be disseminated at the point of publication and used in research and practice. In cases where existing measures are revised with the use of multilevel methods, the revised multilevel dimensions should be published and disseminated by assessment developers. Head Start could incentivize assessment publishers to publish revised versions of assessments using multilevel dimensionality by favoring such instruments and implementing them widely across programs.

Although efforts made by assessment developers and researchers will help provide more appropriate instruments for teachers to use, there is also a need to improve teachers' use of child assessment data to ensure that assessor variance remains minimized moving forward. Improved teacher education would promote an emerging education

workforce with a strong foundation in child-centered assessment and ensure fidelity of implementation with high psychometric validity (Hamilton, Halverson, Jackson, Mandinach, Supovitz & Wayman, 2009). Requiring such course work, or at least a certificate program that includes assessment course work, would provide incoming teachers with an understanding of measurement issues in early childhood, such as assessor variance, and hands-on experience geared towards learning how to best administer such assessments in a child-centered manner. This would ultimately improve teachers' use of child assessment data to improve their practice (Hamilton et al., 2009).

Conclusion

The present research is grounded in the major role that scientifically based assessment plays in our national school readiness policy and practice. This is particularly true for the most underserved, vulnerable prekindergarten children who are most dependent on quality early childhood experiences to advance their school readiness competencies across multiple domains of functioning. Early Childhood teachers and program administrators need quality information about children's functioning across time to guide implementation of curricula, to evaluate children's achievement of these important competencies, and to improve the overall efficacy of preschool programs for children from low-income households. The routine uses of valid multidimensional, teacher-report assessments based on our most advanced psychometric science will make optimal contributions to achieving school readiness for all children. Therefore, we need to critically examine the current most widely-used preschool teacher-report assessments using the most advanced psychometric methods and improve them where necessary.

In the last decade, the most advanced assessment research has demonstrated how assessor variance in teacher-report measures of school-aged children can significantly compromise the validity of school-based, child assessment. These studies have shown the significant impact that assessor variance can have on even the most highly developed teacher-report assessments. Moreover, this research has shown that this threat to validity can be statistically addressed using sophisticated, multilevel factor analysis methods to produce a more child-centered examination of the dimensionality and external validity of teacher-report measures.

The present study was the first to bring this investigation of assessor variance to teacher-report assessment of preschool-aged, Head Start children. This study focused on a highly developed multidimensional, early childhood assessment of preschool children's Approaches to Learning--the LTLS. It applied the most rigorous, multilevel psychometric methods to improve the precision of this multidimensional, teacher-report assessment by determining the level of threat assessor variance posed to the validity of the LTLS and then demonstrating that removing high levels of assessor variance improved its validity.

The results made visible a substantial level of assessor variance evident overall and within every item. The analyses indicated an average of 21% assessor variance among the items on the LTLS, which is over four times the amount of assessor variance (i.e., 5%) that the research literature requires to signal the need to use multilevel statistical methods (Pornprasertmanit et al., 2014). These findings bring into question the scientific integrity of using traditional statistical methods to validate the latent structure of measures like the LTLS. As indicated in the psychometric literature, this level of assessor variance calls for the application of multilevel factor analysis methods to

account for the assessor variance found to provide a more precise *child-centered* assessment of children's Approach to Learning abilities observed in the preschool classroom.

This study demonstrated that accounting for assessor variance using multilevel factor analytic methods refined our understanding of the LTLS's dimensions. Results showed that the difference found between the traditional and the multilevel factor methods was the most severe type of difference that can emerge when making this type of comparison -- a change in *both* the number of dimensions identified and the nature of the dimensions. The number of LTLS dimensions dropped from seven dimensions derived from the traditional analysis to six dimensions resulting from the multilevel analysis. In addition, the multilevel factor solution produced a qualitatively different *Demonstrated Engagement in Learning* dimension.

Compared to the traditional dimension, the multilevel *Demonstrated Engagement in Learning* dimension included more items to more robustly define the nature of children's behaviors that demonstrate engagement in learning. These additional items captured in this multilevel dimension represent important ways that children demonstrate how they are engaged in productive independent activities, communicate choices to adults, willingly participate in challenging activities, and express pride in what their initiative and curiosity produces. This inclusion of additional items that assess children's initiative and curiosity is important because it aligns with how Head Start conceptualizes school readiness. Thus, the multilevel factor analytic approach resulted in better coverage of children's engagement than the traditional factor analysis method.

As a result of these refinements, the multilevel dimensions had greater capacity to explain variance in important external outcomes compared to traditional dimensions. The greatest improvement in predictive ability was seen for the *Demonstrated Engagement in Learning* dimension, which was due to fact that the multilevel version of that dimension was more comprehensive than the corresponding traditional factor. The multilevel *Demonstrated Engagement in Learning* factor improved the prediction of academic outcomes by roughly 45% to 80% over the traditional dimension. The improvement in external validity was most evident for children's vocabulary skills five months later where the multilevel dimension predicted 80% more variance than the traditional dimension. Thus, as expected, using multilevel factor analysis to remove assessor variance and focus on child-level variance resulted in dimensions that better predict children's outcomes.

This study sounds an alarm to alert the early childhood education community to the need to seriously attend to assessor variance and recognize the need to use multilevel statistical methods to reduce its threat to measurement validity (Reise et al., 2005, p. 127; Cronbach, 1976). Cronbach warned that if multilevel statistical methods are not utilized, then "educational research, and a great deal of social science is in serious trouble… [traditional statistical] methods have generated false conclusions in many studies." His prophetic warning, while spurring to action the pioneering work of Raudenbush and Bryk, has been heretofore ignored by the early childhood education community. The present study provides empirical evidence to support Cronbach's charge that we must apply our most advanced multilevel methods to the development and validation of early

childhood teacher-report assessment, especially for prekindergarten programs serving children living in poverty.

Moving forward this can be accomplished by identifying the most widely used teacher-report assessments in Head Start and then applying the multilevel statistical approaches that were used in present study to account for assessor variance and thereby improve the quality of early childhood assessment used in major national programs like Head Start. Two large-scale, national early childhood projects initiated over a decade ago can inform these advances. The National Research Council's systematic review of widely used early childhood assessments provides a model of how to identify widely used early childhood assessments (NRC, 2008). This model could be employed conduct a systematic review to identify which teacher-report assessments are most widely used today in federal or state funded preschool programs for children from low-income households. In addition, the Preschool Curriculum Evaluation Research (PCER) approach, which was used to scientifically test the efficacy of widely used preschool curricula, could serve as a guide for how to apply the most advanced multilevel psychometric methodology to address assessor variance and test the validity of the most widely used teacher-report preschool assessments (PCER, 2008). Guided by the model used in the PCER study, researchers could employ a uniform analytic protocol to use multilevel methods to test for assessor variance and examine the validity of these assessments.

In addition to a national scientific evaluation of the most widely used teacher-report assessments in early childhood, researchers and publishers also need to focus on longer-term efforts to minimize the threat of assessor variance to the validity of teacher-

report assessment in early childhood. This would require test developers of new teacher-report assessments to use these advanced multilevel methods to ensure that these measures meet the latest Standards for Educational and Psychological Testing with specific attention to assessor variance in teacher-report child assessment (AERA et al., 2014). Also, we need to enhance our preparation of early childhood educators to ensure that the emerging, education workforce is knowledgeable about threats of assessor variance to child-centered assessment and that they have received adequate training to administer teacher-report, child assessments with fidelity to ensure the validity of the assessments (Hamilton, Halverson, Jackson, Mandinach, Supovitz & Wayman, 2009).

Clearly, we must move beyond the mere assertion in the Head Start Act requiring the use of only 'scientifically based' measurement, to an accountability practice that ensures that all assessments being used to measure school readiness have demonstrated validity evidencing the application of state-of-the-art scientific methods (NRC, 2008; Head Start Act, P.L. 110-134, 2007). The efficacy of educational programs is in serious jeopardy without the application of these advanced psychometric methods. This study demonstrated that the multilevel methods that remove the threat of assessor variance substantially altered subsequent empirical analysis. Future researchers and practitioners should carefully consider the use of teacher-report measures based on traditional factor analytic methods, especially if data are being used for important decision making with respect to our nation's most vulnerable young children. Our assessment *must be ready* to meet the needs of our most vulnerable young students with the best scientifically based, child-centered information to improve teachers' classroom interventions and increase the

likelihood that the children will be *ready for school* across all relevant domains of functioning.

APPENDIX A: Rotated Factor Loadings for the 5-Factor Solution using the Traditional Approach

TABLE A1

Rotated Factor Loadings for the 5-Factor Solution using the Traditional Approach

	Factor				
Item	1	2	3	4	5
Perseveres with assistance and encouragement	-.02	**.63**	.03	.21	.22
Previous attempts unsuccessful, still tries	-.06	**.66**	.04	.09	.34
Develops plan after considering consequences	.05	**.46**	.38	.17	-.07
Screens out noise and distractions	.21	**.50**	-.06	.31	-.04
Basic understanding of cause and effect	.07	.37	**.41**	.19	-.14
Takes turn in group without reminder	**.60**	.15	.00	.25	-.09
Accepts peer advice by following it	**.77**	.05	.13	-.04	-.08
Plays with child during free play	.30	-.20	.34	.37	.20
Listens and waits for turn to speak	**.68**	.06	-.06	.25	-.07
Self-selects activity without direction	.23	.11	.07	**.48**	.15
Tries new task instead of familiar	-.05	**.57**	.20	.07	.22
Voluntarily demonstrates academic skills	-.12	.04	**.70**	.12	.31
Initiates activity with children	.19	.02	.26	**.43**	.17
Changes strategies when solution not working	-.02	**.55**	.34	.14	.00
Sense of humor with errors	.25	.24	**.45**	-.15	.09
Focused on individual activity, 20 minutes	.04	.14	.10	**.77**	-.04
Attentive when spoken to by teacher	**.70**	.09	.02	.12	.04
Refrains from aggression when frustrated	**.74**	-.02	-.16	.11	.06

TABLE A1 Continued

Rotated Factor Loadings for the 5-Factor Solution using the Traditional Approach

	Factor				
Item	1	2	3	4	5
Responds positively to suggestions for alternate approach	**.62**	.13	.07	.01	.17
Participates in activity or lesson	.32	.12	-.04	.29	**.49**
Receptive when asked to participate in new task	.17	.33	.06	.11	**.50**
Self-corrects errors	.05	**.56**	.29	.13	.00
Willingly participates in unfamiliar activities	.14	.35	.08	.07	**.51**
Responds to questions about ideas without becoming upset	**.52**	.11	.23	-.06	.18
Compares new task to previous re: what worked	.04	.38	**.54**	.09	-.11
Focused on group activity, 10 minutes	.26	.03	.03	**.62**	.20
Perseveres with little input from teacher	.10	**.69**	.07	.16	-.02
Develops plan for multi-step activity	.01	.21	**.65**	.21	-.12
Focused on individual activity, 30 minutes	.05	.16	.11	**.68**	-.10
Works independently with minimal supervision	.24	.37	.07	.39	-.03
Asks teacher for a task	.27	.00	**.47**	-.06	.01
Teaches another child a skill	.09	.18	**.61**	.25	-.11
Works cooperatively to complete task	**.51**	-.04	.23	.29	.05
Asks questions and shares ideas	.13	.07	**.65**	.09	.14
Maintains essential role in small group	.06	.19	**.59**	.24	-.08
Attentive when teacher leads group activity	**.66**	.16	-.07	.21	.00

TABLE A1 Continued

Rotated Factor Loadings for the 5-Factor Solution using the Traditional Approach

	Factor				
Item	1	2	3	4	5
Accepts teacher advice by following it	**.75**	.06	.01	.09	.11
Identifies alternate uses for object	.00	.16	**.50**	.23	.08
Verbalizes frustration and asks for help	.29	-.11	**.60**	-.09	.05
Seeks answers by engaging with materials and people	.24	.13	**.53**	.07	.06
Shows interest and positive attitude toward new activities	.35	.28	.15	-.04	.39
Communicates problems may have more than one solution	.10	**.40**	**.57**	.00	-.09
Focused on individual activity, 10 minutes	.19	.02	.01	**.70**	.15
Tries activity when solution not forthcoming	.16	**.82**	.02	.02	-.02
Perseveres when distracting activities available	.21	**.80**	-.05	.10	-.06
Practices activity without prompting	.06	**.61**	.11	.18	.08
Verbalizes possible consequences	.09	.19	**.65**	.13	-.09
Engages in activity previously challenging	.08	**.77**	.13	-.09	.10
Helps, shares, discusses in group	**.44**	.02	**.40**	.15	.09
Demonstrates pride in work products	-.01	-.08	**.74**	.05	.37
Learns by accepting constructive feedback	**.45**	.22	.22	.04	.13
Verbalizes frustration but continues working	.32	.34	**.43**	-.17	.02
Responds positively to assistance	**.60**	.01	.28	-.09	.11
Guesses even when unsure	.16	.17	**.51**	-.06	.26
Engages in activity without need for approval	.15	**.46**	.10	.16	.24

APPENDIX B: Rotated Factor Loadings for the 6-Factor Solution using the Traditional Approach

TABLE B1

Rotated Factor Loadings for the 6-Factor Solution using the Traditional Approach

	Factor					
Item	1	2	3	4	5	6
Perseveres with assistance and encouragement	.22	.00	.17	.38	.37	-.08
Previous attempts unsuccessful, still tries	.17	-.04	.06	**.49**	.38	-.01
Develops plan after considering consequences	**.67**	.13	.05	.13	.14	-.07
Screens out noise and distractions	.25	.22	.27	.10	.31	-.21
Basic understanding of cause and effect	**.72**	.15	.06	.05	.06	-.08
Takes turn in group without reminder	.23	**.58**	.22	.00	.06	-.12
Accepts peer advice by following it	.14	**.70**	-.03	-.07	.09	.10
Plays with child during free play	.15	.25	.37	.20	-.22	.23
Listens and waits for turn to speak	.09	**.64**	.24	-.03	.06	-.09
Self-selects activity without direction	.20	.22	**.45**	.25	-.06	-.07
Tries new task instead of familiar	.29	-.03	.04	.36	.32	.05
Voluntarily demonstrates academic skills	.34	-.16	.12	.32	-.05	**.47**
Initiates activity with children	.24	.17	**.41**	.23	-.10	.10
Changes strategies when solution not working	**.58**	.03	.05	.19	.25	-.03
Sense of humor with errors	.35	.23	-.17	.14	.14	.27
Focused on individual activity, 20 minutes	.10	-.01	**.80**	-.04	.14	.01
Attentive when spoken to by teacher	.05	**.64**	.13	.07	.10	.04
Refrains from aggression when frustrated	.07	**.74**	.05	.14	-.12	-.15

TABLE B1 Continued

Rotated Factor Loadings for the 6-Factor Solution using the Traditional Approach

	Factor					
Item	1	2	3	4	5	6
Responds positively to suggestions for alternate approach	.14	**.61**	-.04	.25	.02	.04
Participates in activity or lesson	-.13	.29	.30	**.54**	.01	.10
Receptive when asked to participate in new task	.03	.16	.08	**.61**	.11	.10
Self-corrects errors	**.52**	.09	.05	.17	.28	-.04
Willingly participates in unfamiliar activities	.07	.14	.03	**.62**	.11	.10
Responds to questions about ideas without becoming upset	.26	**.52**	-.12	.26	-.03	.13
Compares new task to previous re: what worked	**.69**	.08	-.01	.04	.13	.09
Focused on group activity, 10 minutes	-.07	.20	**.65**	.20	.05	.08
Perseveres with little input from teacher	.22	.06	.20	.08	**.54**	-.01
Develops plan for multi-step activity	**.65**	.02	.17	-.04	.05	.20
Focused on individual activity, 30 minutes	.13	.00	**.72**	-.11	.17	.02
Works independently with minimal supervision	.21	.20	**.40**	.05	.27	-.05
Asks teacher for a task	.13	.17	.02	-.08	.12	**.41**
Teaches another child a skill	**.55**	.06	.25	-.07	.09	.24
Works cooperatively to complete task	.08	**.43**	.34	.02	.02	.20
Asks questions and shares ideas	.38	.08	.10	.15	.01	**.42**
Maintains essential role in small group	**.51**	.02	.24	-.05	.11	.26
Attentive when teacher leads group activity	-.02	**.60**	.23	.02	.20	.00

TABLE B1 Continued

Rotated Factor Loadings for the 6-Factor Solution using the Traditional Approach

	Factor					
Item	1	2	3	4	5	6
Accepts teacher advice by following it	-.14	**.66**	.14	.07	.18	.17
Identifies alternate uses for object	.37	-.03	.23	.12	.06	.25
Verbalizes frustration and asks for help	.08	.16	.04	-.09	.09	**.58**
Seeks answers by engaging with materials and people	.29	.16	.12	.04	.14	.38
Shows interest and positive attitude toward new activities	-.13	.26	.03	**.40**	.28	.31
Communicates problems may have more than one solution	**.56**	.08	-.01	-.02	.27	.23
Focused on individual activity, 10 minutes	-.07	.14	**.74**	.14	.05	.04
Tries activity when solution not forthcoming	.10	.07	.12	.05	**.71**	.06
Perseveres when distracting activities available	.08	.13	.20	.00	**.71**	.00
Practices activity without prompting	.19	.02	.22	.16	**.47**	.04
Verbalizes possible consequences	**.63**	.09	.08	-.01	.04	.24
Engages in activity previously challenging	.19	.03	-.04	.20	**.61**	.10
Helps, shares, discusses in group	.18	.35	.19	.07	.05	.31
Demonstrates pride in work products	.15	-.12	.13	.30	-.02	**.65**
Learns by accepting constructive feedback	.07	.38	.09	.13	.23	.24
Verbalizes frustration but continues working	.13	.19	-.05	-.05	**.42**	**.43**
Responds positively to assistance	-.05	**.49**	-.01	.02	.16	.38
Guesses even when unsure	.18	.09	-.02	.25	.14	**.43**
Engages in activity without need for approval	.03	.09	.21	.30	.37	.14

APPENDIX C: Comparison between the Traditional Factor Structure Reported in McDermott et al., 2011 to the Results using Mplus "EFA" Procedure

McDermott et al., 2011 identified seven dimensions of Approaches to Learning as part of their original validation of the Learning-to-Learn Scales. The factors were named Strategic Planning (e.g., item, "Developed a plan of action after considering the possible consequences"), Effectiveness Motivation ("Voluntarily engages in an activity that has previously posed some challenges"), Interpersonal Responsiveness in Learning ("Responds positively to suggestions for an alternative way to complete a task or activity (i.e., positive verbal or nonverbal response"), Vocal Engagement in Learning ("Voluntarily demonstrates skills and knowledge (e.g., "Listen to me count to 10," "I wrote my name."), Sustained Focus in Learning ("Stays focused on an individual, self-directed activity for more than 10 minutes"), Acceptance of Novelty and Risk ("Acts in a receptive and confident way when asked to participate in a new task or activity"), and Group Learning ("Initiates an appropriate activity with another child or children without direction from teacher or teacher assistant (e.g., building with blocks, starting a puzzle").

Similarly, the Mplus analysis produced 7 dimensions of Approaches to Learning (see Results Chapter). Six of the dimensions were interpreted similarly. The 7-Factor solution represented factors that were consistent with the federal school readiness framework for Head Start and empirical research linking these skills to success in the early classroom (see Administration for Children and Families, 2015a; Hyson, 2008). *Strategic Planning* (Factor 1) most closely aligned with Head Start's subdomain of Cognitive Self-Regulation as it captured children ability to demonstrate flexibility in thinking and behavior (ACF, 2015a, Goal P-ATL 9). *Interpersonal Responsiveness in*

Learning (Factor 2) corresponds to the Emotional & Behavioral Self-Regulation sub-domain (ACF, 2015a, Goal P-ATL 1-4). These dimensions monitor the behavioral demands of responding to classroom routines and interacting appropriately with peers and adults. *Acceptance of Novelty and Risk* (Factor 3) corresponds most closely with the sub-domain of Initiative and Curiosity (ACF, 2015a, Goal P-ATL 10-11). Children developing these skills show an interest and curiosity in their classroom environment. *Sustained Focus in Learning* (Factor 4) most closely aligned with Head Start's subdomain of Cognitive Self-Regulation since it required children to be able to persist in tasks and maintain focus and attention with minimal adult support (ACF, 2015a, Goal P-ATL 6-7). They capture children's flexibility in thinking and ability to control cognitive thought processes. *Effectiveness Motivation* (Factor 5) correspond to Head Start's subdomain of Cognitive Self-Regulation. They capture children's flexibility in thinking and ability to control cognitive thought processes. *Demonstrated Engagement in Learning* (Factor 6) most closely correspond to Head Start's Early Learning Outcomes Framework conceptualization of children's *Initiative & Curiosity*. Children developing these skills show an initiative to engage in their classroom environment. Finally, *Group Learning* (Factor 7) mirrored skills under Head Start's Emotional and Behavioral Self-Regulation subdomain of Approaches to Learning. Specifically, under Goal P-ATL 4 Head Start children are expected to be able to wait for their turn, refrain from aggressive behavior towards other, and began to understand the consequences of behavior. This 7-factor solution best reflected the Head Start framework and existing research on distinct aspects of Approaches to Learning that are predictive of children's academic outcomes.

However, one factor, Demonstrated Engagement in Learning was conceptually similar to McDermott et al. (2011), Vocal Engagement in Learning dimension. One major distinction is that McDermott et al.'s analysis interpreted the dimension as purely vocal expressions of engagement whereas the traditional factor identified in Mplus was interpreted as demonstrations of initiative in the classroom. Various reasons for finding differences between the two methods could be due to the factor rotation options and factor extraction methods that differ between the software packages used to carry out the factor analysis.

To ensure a fair comparison the multilevel and traditional factor analyses, we determined that important factors must be held constant across the multilevel and traditional approaches including the estimator, and the factor rotation procedure (Ford, MacCallum, Tait, 1996; Osborne & Costello, 2009). Rather than using McDermott and colleagues (2011) traditional factor solution which would differ from any multilevel model we would estimate, we carried carry out a new traditional factor analysis using the same estimator and factor rotation procedure. This allowed us to attribute differences in the factor solutions to the traditional versus multilevel approaches rather than these other factors that could influence the final factor solution.

APPENDIX D: Abbreviated Section of the Learning-to-Learn Scales Administration Form Containing the Group Learning Dimension Items

	Consistently Applies	Sometimes Applies	Does Not Apply
27. Actively perseveres with a difficult task with little input from teacher or teacher assistant.	O	O	O
28. Develops a plan for multi-step activity (e.g., "First, I'm going to turn on the oven. Then, I will mix the cake and bake it.").	O	O	O
29. Stays focused on an individual, self-selected activity for more than 30 minutes.	O	O	O
30. Works independently at assigned task with minimal supervision.	O	O	O
31. Asks teacher or teacher assistant for a task to perform or an activity to engage in.	O	O	O
32. Teaches another child a new task or skill.	O	O	O
33. Works cooperatively with another child or small group of children to complete an activity.	O	O	O
34. Willingly asks questions and shares ideas on a variety of topics and tasks.	O	O	O
35. Maintains an essential role when participating in a small group activity (e.g., other children depend on this child for direction).	O	O	O

Note. Group Learning contains item 32, 33, and 35.

BIBLIOGRAPHY

Administration for Children and Families (2015a). Head Start Early Learning Outcomes Framework 2015. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/hs/sr/approach/cdelf

Administration for Children and Families. (2015b). FY 2015 CCDF Allocations. Retrieved from http://www.acf.hhs.gov/programs/occ/resource/fy-2015-ccdf-allocations-including-realloted-funds

Administration for Children and Families (2014). Framework for Effective Practice: Supporting School Readiness for All Children. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/tta-system/teaching/practice.

Administration for Children and Families (2012). Advisory Committee on Head Start Research and Evaluation: Final Report. Retrieved from https://www.acf.hhs.gov/sites/default/files/opre/eval_final.pdf

Administration for Children and Families (2010). Head Start Early Learning Outcomes Framework: Ages Birth to Five. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/hs/sr/approach/pdf/ohs-framework.pdf.

Aitkin, M., Bennett, S. N., & Hesketh, J. (1981). Teaching styles and pupil progress: a re-analysis. *British Journal of Educational Psychology*, *51*(2), 170-186.

Ali, P. A., & Watson, R. (2016). Peer review and the publication process. *Nursing Open*, *3*(4), 193–202. http://doi.org/10.1002/nop2.51

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. Washington, DC: AERA Publications.

Ayduk, O., Mendoza-Denton, R., Mischel, W., Downey, G., Peake, P. K., & Rodriguez, M. (2000). Regulating the interpersonal self: strategic self-regulation for coping with rejection sensitivity. *Journal of personality and social psychology*, *79*(5), 776.

Barghaus, K. M., & Fantuzzo, J. W. (2014). Validation of the Preschool Child Observation Record: Does It Pass the Test for Use in Head Start?. *Early Education and Development*, *25*(8), 1118-1141.

Barghaus, K., LeBoeuf, W., Fantuzzo, J., Brumley, B., Coe, K. (2017). *A Comprehensive Examination of the School District of Philadelphia's Kindergarten Classroom Engagement Scale Technical Report*. Philadelphia, PA: Penn Child Research Center.

Barnett, W. S. (2011). Effectiveness of early educational intervention. *Science*,*333*(6045), 975-978.

Barnett, W. S., Weisenfeld, G. G., Brown, K., Squires, J., & Horowitz, M. (2016). Implementing 15 Essential Elements for High Quality: A State and Local Policy Scan. *National Institute for Early Education Research*.

Bennett N. 1976. *Teaching Styles and Pupil* Cohen P. 1998. Black concentration effects on *Progress.* London: Open Books

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., ... & Gill, S. (2008). Promoting academic and social-emotional school readiness: The Head Start REDI program. *Child development*, *79*(6), 1802-1817.

Blair, C., & Diamond, A. (2008). Biological processes in prevention and intervention: The promotion of self-regulation as a means of preventing school failure. *Development and Psychopathology*, *20*(03), 899-911.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*(2), 647-663.

Bodovski, K. & Farkas, G. Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *The Elementary School Journal*. 108(2), 115-130.

Bronfenbrenner, U. (1995). Developmental ecology through space and time: A future perspective. In P. Moen, G. H. Elder, Jr., & K. Luscher (Eds.), *Examining lives in context: Perspectives on the ecology of human development* (pp. *619*-647). Washington, DC: American Psychological Association.

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. *Handbook of child psychology*.

Bulotsky-Shearer, R. J., Bell, E. R., Romero, S. L., & Carter, T. M. (2012). Preschool interactive peer play mediates problem behavior and learning for low-income children. *Journal of Applied Developmental Psychology*, *33*(1), 53-65.

Bulotsky-Shearer, R. J., Manz, P. H., Mendez, J. L., McWayne, C. M., Sekino, Y., & Fantuzzo, J. W. (2012). Peer play interactions and readiness to learn: A protective influence for African American preschool children from low-income households. *Child Development Perspectives*, *6*(3), 225-231.

Bustamante, A. S., White, L. J., & Greenfield, D. B. (2017). Approaches to learning and school readiness in Head Start: Applications to preschool science. *Learning and Individual Differences*, *56*, 112-118.

Cabell, S. Q., Justice, L. M., McGinty, A. S., DeCoster, J., & Forston, L. D. (2015). Teacher–child conversations in preschool classrooms: Contributions to children's vocabulary development. *Early Childhood Research Quarterly*, *30*, 80-92.

Calkins, S. D., & Fox, N. A. (2002). Self-regulatory processes in early personality development: A multilevel approach to the study of childhood social withdrawal and aggression. *Development and psychopathology*, *14*(03), 477-498.

Camilli, G. (2006). Test fairness. *Educational measurement*, *4*, 221-256.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. *Handbook of research on teaching*, 171-246.

Carlson, A. G. (2014). *Kindergarten fine motor skills and executive function: Two non-academic predictors of academic achievement* (Doctoral dissertation). Retrieved from ProQuest Information & Learning (US).

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*,*80*(1), 219-251.

Child Trends. (2010). A review of school readiness practices in the states. Retrieved from http://www.childtrends.org/wp-content/uploads/2013/05/2010-14-SchoolReadinessStates.pdf.

Child Care and Development Block Grant Act of 2014, S. 1086, 113th Cong. (2014). Retrieved from https://www.acf.hhs.gov/sites/default/files/occ/child_care_and_development_block_grant_markup.pdf.

Cicirelli, V. G. (1969). Project Head Start, a national evaluation: Summary of the study. *Britannica Review of American Education*, *1*.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155.

Coleman, J. S. (1966). Equality of Educational Opportunity (COLEMAN) Study (EEOS). ICPSR06389-v3. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2007-04-27. http://doi.org/10.3886/ICPSR06389.v3

Cooper, H. M. & Hedges, L. V. (Eds.) (1994). The handbook of research synthesis. New York: The Russell Sage Foundation.

Cunha, F., Heckman, J. J., Lochner, L., & Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education*, *1*, 697-812.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental psychology*, *33*(6), 934.

Coolahan, K., Fantuzzo, J., Mendez, J., & McDermott, P. (2000). Preschool peer interactions and readiness to learn: Relationships between classroom peer play and learning behaviors and conduct. *Journal of Educational Psychology*, *92*(3), 458.

Cronbach, L. J. (1976). Research on classrooms and schools: Formulation of questions, design, and analysis. Unpublished manuscript, Stanford University, Stanford Evaluation Consortium, School of Education.

D'Haenens, E., Van Damme, J., & Onghena, P. (2010). Multilevel exploratory factor analysis: illustrating its surplus value in educational effectiveness research. *School Effectiveness and School Improvement*, *21*(2), 209-235.

DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of item candidates: the PROMIS qualitative item review. *Medical care*, *45*(1), S12.

Dickinson, D. K., & Porche, M. V. (2011). Relation between language experiences in preschool classrooms and children's kindergarten and fourth-grade language and reading abilities. *Child Development*, *82*(3), 870-886.

DiPerna, J. C., Lei, P. W., & Reid, E. E. (2007). Kindergarten predictors of mathematical growth in the primary grades: An investigation using the Early Childhood Longitudinal Study--Kindergarten cohort. *Journal of Educational Psychology*, *99*(2), 369.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1-11.

Dobbs, J., Doctoroff, G. L., Fisher, P. H., & Arnold, D. H. (2006). The association between preschool children's socio-emotional functioning and their mathematical skills. *Journal of Applied Developmental Psychology*, *27*(2), 97-108.

Dobbs-Oates, J., & Robinson, C. (2012). Preschoolers' mathematics skills and behavior: Analysis of a national sample. *School Psychology Review*, *41*(4), 371.

Downer, J. T., Booren, L. M., Lima, O. K., Luckner, A. E., & Pianta, R. C. (2010). The Individualized Classroom Assessment Scoring System (inCLASS): Preliminary reliability and validity of a system for observing preschoolers' competence in classroom interactions. *Early childhood research quarterly*, *25*(1), 1-16.

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830-837.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, *10*(1), 61-82.

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

Duncan, G. J., & Magnuson, K. (2011). The nature and impact of early achievement skills, attention skills, and behavior problems. *Whither opportunity*, 47-70.

Dunn, E. C., Masyn, K. E., Jones, S. M., Subramanian, S. V., & Koenen, K. C. (2015a). Measuring Psychosocial Environments Using Individual Responses: an Application of Multilevel Factor Analysis to Examining Students in Schools. *Prevention Science*, 1-16.

Dunn, E. C., Masyn, K. E., Johnston, W. R., & Subramanian, S. V. (2015b). Modeling contextual effects using individual-level data and without aggregation: an illustration of multilevel factor analysis (MLFA) with collective efficacy. *Population Health Metrics*, *13*(1), 12

Edmunds, J. M., Beidas, R. S., & Kendall, P. C. (2013). Dissemination and implementation of evidence–based practices: Training and consultation as implementation strategies. *Clinical Psychology: Science and Practice*, *20*(2), 152-165.

Eisenberg, N., & Spinrad, T. L. (2004). Emotion-related regulation: Sharpening the definition. *Child development*, 334-339.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Evans, G. W., & Kim, P. (2013). Childhood poverty, chronic stress, self-regulation, and coping. *Child Development Perspectives*, *7*(1), 43-48.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, *4*(3), 272.

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

Fantuzzo, J. W., Gadsden, V. L., & McDermott, P. A. (2011). An integrated curriculum to improve mathematics, language, and literacy for Head Start children. *American Educational Research Journal*, *48*(3), 763-793.

Fantuzzo, J., LeBoeuf, W., Rouse, H., & Chen, C. C. (2012). Academic achievement of African American boys: A city-wide, community-based investigation of risk and resilience. *Journal of School Psychology*, *50*(5), 559-579.

Fantuzzo, J., Perry, M. A., & McDermott, P. (2004). Preschool approaches to learning and their relationship to other relevant classroom competencies for low-income children. *School Psychology Quarterly*, *19*(3), 212.

Finch, W. H. (2011). A comparison of factor rotation methods for dichotomous data. *Journal of Modern Applied Statistical Methods*, *10*(2), 14.

Forman, S. G., Shapiro, E. S., Codding, R. S., Gonzales, J. E., Reddy, L. A., Rosenfield, S. A., ... & Stoiber, K. C. (2013). Implementation science and school psychology. *School Psychology Quarterly*, *28*(2), 77.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of personality and social psychology*, *78*(2), 350.

Friedman-Krauss, A. H., Barnett, W. S., Weisenfeld, G.G. Kasmin, R., DiCrecchio, N., & Horowitz, M. (2018). The state of preschool 2017. National Institute of Early Education Research. New Brunswick, NJ: Rutgers Graduate School of Education.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American educational research journal*, *38*(4), 915-945.

Gershoff, E. T., Aber, J. L., Raver, C. C., & Lennon, M. C. (2007). Income is not enough: Incorporating material hardship into models of income associations with parenting and child development. *Child development*, *78*(1), 70-95.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., ... & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4-19.

Goldstein, H. (1995). Hierarchical data modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, *20*(2), 201-204.

Goldstein, J., & McCoach, D. B. (2011). The starting line: Developing a structure for

teacher ratings of students' skills at kindergarten entry. *Early Childhood Research & Practice*, *13*(2). Retrieved from http:// les.eric.ed.gov/ fulltext/EJ956366.pdf.

Gorsuch, R. L. (2003). Factor analysis. In J. A Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2, pp. 143-164). Hoboken, NJ: John Wiley.

Graziano, P. A., Reavis, R. D., Keane, S. P., & Calkins, S. D. (2007). The role of emotion regulation in children's early academic success. *Journal of school psychology*, *45*(1), 3-19.

Guo, G., & Zhao, H. (2000). Multilevel modeling for binary data. *Annual review of sociology*, *26*(1), 441-462.

Guskey, T. R. (2003). What makes professional development effective?. *Phi delta kappan*, *84*(10), 748-750.

Halle, T. G., & Darling-Churchill, K. E. (2016). Review of measures of social and emotional development. *Journal of Applied Developmental Psychology*, *45*, 8-18.

Halle, T., Zaslow, M., Wessel, J., Moodie, S., & Darling-Churchill, K. (2011). Understanding and Choosing Assessments and Developmental Screeners for Young Children Ages 3-5: Profiles of Selected Measures. OPRE Report# 2011-23. *Administration for Children & Families*.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., Wayman, J. C., ... & Steele, J. L. (2009). Using student achievement data to support instructional decision making.

Hammond, H., Skidmore, L., Wilcox-Herzog, A. & Kaufman, J. (2013). Creativity and creativity programs. *International guide to student achievement*, 292-295.

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist*, *27*(3), 353-383.

Head Start Research and Evaluation Advisory Committee. (2012). Advisory committee on Head Start research and evaluation final report. Washington, DC: U.S. Department of Health and Human Services.

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

Hendrick, J. (2001). The whole child: Developmental education for the early years. Prentice Hall.

Holmes, R. M., Romeo, L., Ciraola, S., & Grushko, M. (2015). The relationship between creativity, social play, and children's language abilities. *Early Child Development and Care*, *185*(7), 1180-1197.

Howard, E., Fantuzzo, J., Flanagan, K., Williams, R., Tucker, N., Feng, L., . . . Brumley, B. (2016). *Pennsylvania 2015 Kindergarten Entry Inventory (KEI) technical memorandum of findings.* (Research Report 2015 Child Cohort). Washington, DC: American Institutes for Research.

Howse, R. B., Calkins, S. D., Anastopoulos, A. D., Keane, S. P., & Shelton, T. L. (2003). Regulatory contributors to children's kindergarten achievement. Early Education and Development, 14, 101–120. http://dx.doi.org/10.1207/s15566935eed1401_7.

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. 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.

Hurd, M. D., McFadden, D., Chand, H., Gan, L., Menill, A., & Roberts, M. (1998). Consumption and savings balances of the elderly: Experimental evidence on survey response bias. *Frontiers in the Economics of Aging*, 353-392.

Hyson, M. (2008). *Enthusiastic and engaged learners: Approaches to learning in the early childhood classroom*. New York: Teachers College Press.

Isaacs, J. B. & Brookings. (2011). The recession's ongoing impact on America's children: Indicators of children's economic well-being through 2011.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *166*(3), 279-300.

Jirout, J., & Klahr, D. (2012). Children's scientific curiosity: In search of an operational definition of an elusive concept. *Developmental Review*, *32*(2), 125-160.

Kagan, S. L., Moore, E., & Bredekamp, S. (1995). *Reconsidering children's early development and learning: Toward common views and vocabulary*. Washington, DC: National Education Goals Panel.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31-36.

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

Kaplan, M., & Mead, S. (2017). *The Best Teachers for Our Littlest Learners? Lessons from Head Start's Last Decade*. Retrieved from https://bellwethereducation.org/publication/best-teachers-our-littlest-learners-lessons-head-start%E2%80%99s-last-decade.

Karoly, L. A., Kilburn, R. M., & Cannon, J. S. (2005). *Early childhood interventions: Proven results, future promise.* Retrieved from the RAND Corporation website: http://www. rand.org/pubs/monographs/2005/RAND_MG341.pdf

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment* (Vol. 53). John Wiley & Sons.

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *The Annals of Family Medicine*, *2*(3), 204-208.

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate behavioral research*, *51*(6), 881-898.

Konold, T., & Cornell, D. (2015). Multilevel multitrait–multimethod latent analysis of structurally different and interchangeable raters of school climate. *Psychological assessment*, *27*(3), 1097.

Li-Grining, C. P., Votruba-Drzal, E., Maldonado-Carreño, C., & Haas, K. (2010). Children's early approaches to learning and academic trajectories through fifth grade. *Developmental Psychology*, *46*(5), 1062–1077. http://doi.org/10.1037/a0020066

Little, J. (2013). Multilevel confirmatory ordinal factor analysis of the Life Skills Profile–16. *Psychological assessment*, *25*(3), 810.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304-316.

Maier, M. F. (2011). *Examining preschoolers' trajectories of individual learning behaviors: The influence of approaches to learning on school readiness* (Doctoral dissertation). Retrieved from ProQuest Information & Learning (US).

Marcon, R. A. (2002). Moving up the Grades: Relationship between Preschool Model and Later School Success. *Early Childhood Research & Practice*, *4*(1).

Mashburn, A. J. (2014). The importance of quality prekindergarten programs for promoting school readiness skills. *Wellbeing*.

Mathematica Policy Research. (2007). *Measuring children's progress from preschool through third grade*. Princeton, NJ: Author.

Matthews, J. S., Kizzie, K. T., Rowley, S. J., & Cortina, K. (2010). African Americans and boys: Understanding the literacy gap, tracing academic trajectories, and evaluating the role of learning-related skills. *Journal of Educational Psychology*, *102*(3), 757.

McClelland, M. M., Acock, A. C., & Morrison, F. J. (2006). The impact of kindergarten learning-related skills on academic trajectories at the end of elementary school. *Early Childhood Research Quarterly*, *21*(4), 471-490.

McClelland, M. M., Acock, A. C., Piccinin, A., Rhea, S. A., & Stallings, M. C. (2013). Relations between preschool attention span-persistence and age 25 educational outcomes. *Early Childhood Research Quarterly*, *28*(2), 314-324.

McClelland, M. M., Morrison, F. J., & Holmes, D. L. (2000). Children at risk for early academic problems: The role of learning-related social skills. *Early Childhood Research Quarterly*, *15*(3), 307-329.

McClelland, M. M., & Morrison, F. J. (2003). The emergence of learning-related social skills in preschool children. *Early Childhood Research Quarterly*, *18*(2), 206-224.

McDermott, P. A., Fantuzzo, J. W., Warley, H. P., Waterman, C., Angelo, L. E., Gadsden, V. L., & Sekino, Y. (2011). Multidimensionality of teachers' graded responses for preschoolers' stylistic learning behavior: The Learning-to-Learn Scales. *Educational and Psychological Measurement*, *71*(1), 148-169.

McDermott, P. A., Fantuzzo, J. W., Waterman, C., Angelo, L. E., Warley, H. P., Gadsden, V. L., & Zhang, X. (2009). Measuring preschool cognitive growth while it's still happening: The Learning Express. *Journal of school psychology*,*47*(5), 337-366.

McDermott, P. A., Green, L. F., Francis, J. M., & Stott, D. H. (2000). Preschool learning behaviors scale. *Philadelphia, PA: Edumetric and Clinical Science*.

McDermott, P. A., Leigh, N. M., & Perry, M. A. (2002). Development and validation of the Preschool Learning Behaviors Scale. *Psychology in the Schools*, *39*(4), 353-365.

McDermott, P. A., Rikoon, S. H., & Fantuzzo, J. W. (2014). Tracing children's approaches to learning through Head Start, kindergarten, and first grade: Different pathways to different outcomes. *Journal of Educational Psychology*, *106*(1), 200.

McDermott, P. A., & Watkins, M. W. (1987). Microcomputer systems manual for McDermott Multidimensional Assessment of Children (IBM version). San Antonio, TX: Psychological Corporation

Mead, S. (2014). Renewing Head Start's promise: Invest in what works for disadvantaged preschoolers. *Washington, DC: Bellwether Education Partners and Results. Retrieved from http://bellwethereducation.org/publication/RenewingHeadStartsPromise*.

Meisels, S. J., & Atkins-Burnett, S. (2004). The Head Start National Reporting System. *Young Children*, *59*(1), 64-66.

Menaker, M. R., Steinberg, C. M., Angelo, L. E., & McDermott, P. A. (2002, June). Forging the link between children's teachable learning behaviors and the Head Start curriculum. Poster presented at the Head Start 6th National Research Conference, Washington, DC.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 4, 525–543.

Mischel, W., Shoda, Y., & Rodriguez, M. I. (1989). Delay of gratification in children. *Science*, *244*(4907), 933-938.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H.,et al., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, *108*(7), 2693-2698.

Muraki, E., & Bock, D. (2003). PARSCALE for Windows. *Chicago: Scientific Software International*.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*(4), 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376-398.

Muthén, B.O., & Sattora, A. (1995). Technical aspects of Muthen's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika*, 60, 489-503.

National Education Goals Panel. (1995). National education goals report executive summary: Improving education through family-school-community partnerships. National Education Goals Panel, Washington, DC.

National Research Council. (2008). *Early Childhood Assessment: Why, What, and How.* Committee on Developmental Outcomes and Assessments for Young Children, C.E. Snow and S.B. Van Hemel, *Editors.* Board on Children, Youth, and Families, Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Office of Management and Budget. (2015).Government performance and results act (GRPA) related materials. https://www.whitehouse.gov/omb/mgmt-gpra/index-gpra.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied measurement in education*, *16*(3), 223-243.

Peters, C., Algina, J., Smith, S. W., & Daunic, A. P. (2012). Factorial validity of the behavior rating inventory of executive function (BRIEF)-teacher form. *Child Neuropsychology*, *18*(2), 168-181.

Phillips, D., Austin, L. J., & Whitebook, M. (2016). The Early Care and Education Workforce. *The Future of Children*, *26*(2), 139-158.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, *88*(5), 879.

Ponitz, C. C., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. *Developmental psychology*, *45*(3), 605.

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research*, *49*(6), 518-543.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P. et al. & Spier, E. (2012). Head Start Impact Study. Final Report. *Administration for Children & Families*.

Ramey, C. T., & Ramey, S. L. (2004). Early learning and school readiness: Can early intervention make a difference?. *Merrill-Palmer Quarterly*, *50*(4), 471-491.

Rampey, B. D., Dion, G. S., & Donahue, P. L. (2009). NAEP 2008: Trends in Academic Progress. NCES 2009-479. *National Center for Education Statistics*.

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

Raver, C. C., Garner, P. W., & Smith-Donald, R. (2007). The roles of emotion regulation and emotion knowledge for children's academic readiness: Are the links causal?. In Pianta, Robert C. (Ed); Cox, Martha J. (Ed); Snow, Kyle L. (Ed), (2007). School readiness and the transition to kindergarten in the era of accountability. , (pp. 121-147). Baltimore, MD, US: Paul H Brookes Publishing, xx, 364 pp.

Razza, R. A., Martin, A., & Brooks-Gunn, J. (2015). Are approaches to learning in kindergarten associated with academic and social competence similarly?. *Child & Youth Care Forum.*

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. *Whither opportunity*, 91-116.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of personality assessment*,*84*(2), 126-136.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological assessment*, *12*(3), 287.

Rikoon, S. H., McDermott, P. A., & Fantuzzo, J. W. (2012). Approaches to learning among Head Start alumni: Structure and validity of the Learning Behaviors Scale. *School Psychology Review*, *41*(3), 272.

Rowand, C., Sprachman, S., Wallace, I., Rhodes, H., and Avellar, H. (2005). *Factors contributing to assessment burden in preschoolers*. Paper presented at the American Association for Public Opinion Research, May, Miami, FL.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, *16*(4), 583-601.

Schriesheim, C.A. & Denisi, A.S. (1980). Item presentation as an influence on questionnaire validity: A field experiment, *Educational and Psychological Measurement*, *40*(1), 175–182.

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys implications for policy and practice. *Educational Evaluation and Policy Analysis*, *36*(3), 259-280.

Schweinhart, L. J., & Weikart, D. P. (1998). Why curriculum matters in early childhood education. *Educational Leadership*, *55*, 57-61.

Serna, L., Nielsen, E., Mattern, N., & Forness, S. R. (2002). Use of different measures to identify preschoolers at-risk for emotional or behavioral disorders: Impact on gender and ethnicity. *Education & Treatment of Children*, *25*(4), 415-437.

Sektnan, M., McClelland, M. M., Acock, A., & Morrison, F. J. (2010). Relations between early family risk, children's behavioral regulation, and academic achievement. *Early Childhood Research Quarterly*, *25*(4), 464-479.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications.

Shoda, Y., Mischel, W., & Peake, P. K. (1990). Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental psychology*, *26*(6), 978.

Shonkoff, J. P. & Phillips, D. A., (Eds.). (2000). *From Neurons to Neighborhoods:: The Science of Early Childhood Development*. National Academies Press.

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2018). How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses. *Annual review of psychology*.

Sprachman, S., Atkins-Burnett, S., Glazerman, S., Avellar, S., and Loewenberg, M. (2007). *Minimizing assessment burden on preschool children: Balancing burden and reliability.* Paper presented at the Joint Statistical Meetings, September, Salt Lake City, UT.

Stapleton, L. M., McNeish, D. M., & Yang, J. S. (2016). Multilevel and single-level models for measured and latent variables when data are clustered. *Educational Psychologist*, *51*(3-4), 317-330.

Stirman, S. W., Gutner, C. A., Langdon, K., & Graham, J. R. (2016). Bridging the gap between research and practice in mental health service settings: An overview of developments in implementation theory and research. *Behavior therapy*, *47*(6), 920-936.

Tabachnick, B. G., & Fidell, L. S. (2001). Using multivariate statistics.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ,: Lawrence Erlbaum Associates, Inc.

Thompson, B. (2002). *Score reliability: Contemporary thinking on reliability issues*. Sage publications.

Trentacosta, C. J., & Izard, C. E. (2007). Kindergarten children's emotion competence as a predictor of their academic competence in first grade.*Emotion*, *7*(1), 77.

Tudge, J. R., Mokrova, I., Hatfield, B. E., & Karnik, R. B. (2009). Uses and misuses of Bronfenbrenner's bioecological theory of human development.*Journal of Family Theory & Review*, *1*(4), 198-210.

Tudge, J. R., Payir, A., Merçon-Vargas, E., Cao, H., Liang, Y., Li, J., & O'Brien, L. (2016). Still misused after all these years? A reevaluation of the uses of Bronfenbrenner's bioecological theory of human development. *Journal of Family Theory & Review*, *8*(4), 427-445.

U.S. Department of Education. (2002). Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for Kindergarten Through First Grade, NCES 2002–05, by Donald A. Rock and Judith M. Pollack, Educational Testing Service, Elvira Germino Hausken, project officer. Washington, DC: 2002.

U. S. Department of Health and Human Services. (2001). *Screening and assessment in head start*. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/tta-system/teaching/eecd/assessment/screening/screeningandass.htm.

U. S. Department of Health and Human Services. (2008). Head start family income guidelines for 2008 (ACF-IM-HS-08-05). Washington, DC: Administration for Children and Families, Office of Head Start.

U. S. Department of Health and Human Services. (2010). *The Head Start child development and early learning framework*. Retrieved from https://eclkc.ohs.acf.hhs.gov/hslc/sr/approach/pdf/OHSApproach-to-School-Readiness_Early-Learning-Framework.pdf.

U. S. Government Accountability Office. (1997). Head Start: Research Provides Little Information on Impact of Current Program. Retrieved from http://www.gao.gov/products/HEHS-97-59.

Vallotton, C., & Ayoub, C. (2011). Use your words: The role of language in the development of toddlers' self-regulation. *Early Childhood Research Quarterly*, *26*(2), 169-181.

Vanneman, A., Hamilton, L., Anderson, J. B., & Rahman, T. (2009). Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress. Statistical Analysis Report. NCES 2009-455. *National Center for Education Statistics*.

Velicer, W. F., & Fava, J. L. (1998). The effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3*, 231–251.

Waber, D. P., Gerber, E. B., Turcios, V. Y., Wagner, E. R., & Forbes, P. W. (2006). Executive functions and performance on high-stakes testing in children from urban schools. *Developmental Neuropsychology*, *29*(3), 459-477.

Waterman, C., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2012). The matter of assessor variance in early childhood education—Or whose score is it anyway?. *Early Childhood Research Quarterly*, *27*(1), 46-54.

Whorrall, J., & Cabell, S. Q. (2016). Supporting children's oral language development in the preschool classroom. *Early Childhood Education Journal*, *44*(4), 335-341.

What Works Clearinghouse. (2015). Head Start. Retrieved from http://ies.ed.gov/ncee/wwc/interventionreport.aspx?sid=636

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under-and overextraction on principal axis factor analysis with varimax rotation. *Psychological methods*, *1*(4), 354.

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. SUNY Press.

Zelazo, P. D., & Carlson, S. M. (2012). Hot and cool executive function in childhood and adolescence: Development and plasticity. *Child Development Perspectives*, *6*(4), 354-360.

Zigler, E., & Styfco, S. J. (2010). *The hidden history of Head Start*. Oxford University Press, USA

Zill, N., & West, J. (2001). Entering Kindergarten: A Portrait of American Children When They Begin School. Findings from the Condition of Education.