2020-12-08

Factor Structure of the Jordan Performance Appraisal System: A Multilevel Multigroup Study Using Categorical and Count Data

Holly Lee Allen
Brigham Young University

Factor Structure of the Jordan Performance Appraisal System: A Multilevel Multigroup Study Using Categorical and Count Data

Holly Lee Allen

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Richard R Sudweeks, Chair
Joseph A. Olsen
Ross A. A. Larsen
Jennifer J. Wimmer

Educational Inquiry, Measurement, and Evaluation

Brigham Young University

ABSTRACT

Factor Structure of the Jordan Performance Appraisal System: A Multilevel Multigroup Study Using Categorical and Count Data

Holly Lee Allen
Educational Inquiry, Measurement, and Evaluation, BYU
Doctor of Philosophy

Development of the Jordan Performance Appraisal System (JPAS) was completed in 1996. This study examined the factor structure of the classroom observation instrument used in the JPAS. Using observed classroom instructional quality ratings of 1220 elementary teachers of Grades 1-6 in the Jordan School District, this study estimated the factor structure of the data and the rater effect on relevant structural parameters. This study also tested for measurement invariance at the within and between levels across teachers of two grade-level groups (a) lower grades: Grades 1-3 and (b) upper grades: Grades 4-6. Factor structure was estimated using complex exploratory factor analysis (EFA) conducted on a subset of the original data. The analysis provided evidence of a three-factor model for the combined groups. The results of multiple confirmatory factor analyses (CFA) conducted using a different subset of the data cross-validated EFA results. Results from multilevel confirmatory factor analysis (MCFA) indicated the three-factor model fit best at both the within and the between levels, and that the intraclass correlation (ICC) was high (.699), indicating significant rater-level variance. Results from a multilevel multigroup confirmatory factor analysis (MLMG-CFA) indicated that the ICC was not significantly different between groups. Results also indicated configural, metric (weak factorial), and scalar (strong factorial) equivalence between groups. This study provided one of the first examples of how to estimate the impact of cluster-level variables such as rater on grouping variables nested at the within level. It provided an example of how to conduct a multilevel multigroup analysis on count data. It also disproved the assumption that counting classroom teaching behaviors was less subjective than using a categorical rating scale. These results will provide substantial information for future developments made to the classroom observation instrument used in the JPAS.

Keywords: classroom teaching observation techniques, factor analysis, structural equation modeling, multilevel multigroup modeling, negative binomial, Poisson

ACKNOWLEDGMENTS

To my parents and siblings, both those I was born to and those I married into, family is everything. You have all been invaluable in your encouragement.

To my sweetheart, Scott who said “Just get it done!” But also provided the means and encouragement by which I could.

To my two daughters. May you always remember to take the best path even when it’s hard. You’ll learn a lot, and ultimately you will change the world with your understanding of it.

To my committee members and coworkers, thank you for your time and your support: To Dr. Joseph Olson and Dr. Richard Sudweeks whose office spaces knew me too well. To Dr. Ross Larsen for teaching me about count data. To Dr. Jenni Wimmer for reminding me to look through my teacher eyes. To Dr. Bryant Jensen for your expertise in measures of classroom teaching quality. To Ben Jameson for your insight and example. To Rebecca Lee for your friendship and expertise on the JPAS. To Dr. Shelly Nordick for being an exceptional mentor and cheerleader.

To Dr. Tenko Raykov, a scholar and friend without whom this analysis would have seemed impossible.

To my fifth-grade teacher, the late James Gooding—a navy man and boxer; an artist; a storyteller; and a master teacher who showed me that teaching is a passion. A calling. An art form so complex that its true effect on individual lives and the world is immeasurable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

In 1983, the publication of *A Nation at Risk: The Imperative for Educational Reform* effected key changes in standards and expectations related to classroom instruction. It identified a "need to improve teaching and learning," calling for "reform and excellence throughout education" (National Commission on Excellence in Education, 1983, p. 5). In 1987, when the National Board for Professional Teaching Standards (NBPTS) was formed, key goals of the NBPTS included maintaining "high and rigorous standards for what accomplished teachers should know and be able to do." As noted by Darling-Hammond (1996), within a decade of these changes, policy makers began narrowly defining teaching quality as "a set of uniform teaching behaviors (often trivial but easy to measure) such as 'keeps a brisk pace of instruction,' 'manages routines,' and 'writes behavioral objectives' with no regard to subject matter, curriculum, or learning" (p. 19). These changes, she noted, had resulted in "promoting teaching that is insensitive to learning while undermining good teaching" (Darling-Hammond, 1996, p. 20). During the 1990s, the developers of Danielson's Framework for Teaching (FfT; Danielson, 1996) and the developers of the Jordan Performance Appraisal System (JPAS; Jordan School District, 1993) focused on these easy to measure, low-inference teaching behaviors. While the FfT was developed for teacher preparation, the JPAS was developed as a formative and summative measure to identify stronger and weaker classroom teaching within a specific, local population that was, at the time, mainly composed of white, middle-class students. Of the two, the JPAS underwent more empirical scrutiny with evidence of multiple principal components analyses (PCA) as an integral part of iterative development. It was used as a formative and summative tool in a handful of school districts across two states. The FfT began to be used in a

significant number of school districts across the United States among more diverse populations and has become mandated in some states as the classroom observation instrument that must be used in evaluating classroom teaching quality.

During the early 2000s classroom observation frameworks introduced elements of classroom instruction where the individual behaviors and needs of students began to emerge alongside low-inference teaching behaviors. The second iteration of the FfT included phrases that indicated attention to diverse learners and a focus on individual needs within the overall framework in addition to having sections that focused on student behaviors as indicators for teaching quality (Danielson, 2007).

The Classroom Assessment Scoring System (CLASS; Pianta et al., 2006) was developed for a different purpose than the FfT or the JPAS. It was developed initially for research purposes and underwent a level of empirical scrutiny that included construct validity, rater agreement, variation in scores by lesson, and variation in scores by grade level. It focused on both teacher and student behaviors, but instead of a general instrument for all grades, the CLASS outlined different behaviors expected of classroom teaching quality dependent on the grade level of the students. Both the Protocol for Language Arts Teaching Observations (PLATO; Grossman & McDonald, 2008) and the Mathematical Quality of Instruction (MQI; Hill et al., 2008) established the role of content area in determining the behaviors that would best serve as indicators of high quality instruction.

While these promising developments did occur in the early 2000s, the No Child Left Behind Act (NCLB, 2001) focused on student outcomes as measured by standardized end-of-year assessments, and as a result, by 2015 statistics indicated that 42 states required student growth as a portion of teacher evaluation and 17 states required student growth to be the

"preponderant criterion in teacher evaluations" (Dorety & Jacobs, 2015, p. iii). This shift placed the study of classroom teaching quality as a component of teaching effectiveness for much of the classroom observation research. Ratings from observation instruments became a means to compare teaching ratings to student achievement as measured by models such as the Tennessee Value Added System (TVAAS; Sanders & Horn, 1994), and student growth percentiles (SGP; Betebenner, 2009). These studies are what Jensen et al. (2019) refer to as teaching effectiveness studies, and they have continued to be of interest to many researchers (Gill et al., 2016; Charalambous et al., 2019) years after the Every Student Succeeds Act (ESSA, 2015) removed student growth measures from the list of educator evaluation requirements, and the American Educational Research Association (AERA) recommended that the limitations of value-added models (VAMs) and other growth models be seriously considered before their inclusion in educator evaluation (AERA, 2015). Other researchers noted that while showing more stability than VAMs, classroom observational measures, which were recommended to replace VAMs, still had lower stability than some measures found in higher education (Polikoff, 2015).

Accounting for Outside Variables

Researchers' increased scrutiny of classroom observation instruments, their use, and their relationship to results from measures of student achievement had a positive impact on the complexity and sophistication of research questions and methods used to answer them. Goe et al. (2008) explain, "The degree to which observations can or should be used for specific purposes depends on the instrument, how that instrument was developed, the level of training and monitoring raters receive, and the psychometric properties of the instrument" (p. 20). Analyzing the psychometric properties of classroom observation instruments includes more

than merely identifying basic statistics or even factor structure. Most pertinent seems to be the capacity of this exploration to uncover the way in which student, classroom, and school variables might impact not only latent trait estimates, but the relationship between behavioral indicators and these latent traits. As Cohen and Goldhaber (2016) observe, "Part of the challenge is that instructional quality is inherently situated. Good teaching likely varies in response to contextual factors" (p. 1). These contextual factors include principal raters, content area, grade level, student demographics, and other variables that might influence ratings of observed classroom teaching quality. The effects of these contextual factors have been examined on numerous occasions. Studies on the impact of time of day (Curby et al., 2011), lesson type (Mikeska et al., 2019; Qi et al., 2014) and rater effect (Casabianca et al., 2015; Gitomer et al., 2014; Jensen et al., 2019), along with studies on student socio-economic, cultural, and linguistic characteristics (Gill et al., 2016; Jensen et al., 2019) indicated that contextual variables impacted ratings of observed classroom teaching quality. Most of these contextual variables were compared using *t*-tests, ANOVA, Multiple Indicators Multiple Causes (MIMIC) models, or other statistical comparisons related to the mean and variance of latent traits or behavioral indicators. These types of comparisons—while important in the general sense of understanding the impact of contextual variables—do not examine the structural level of this impact. The structural influence of contextual variables drives the theories behind testing measurement invariance. Invariance testing conducted within a Structural Equation Modeling (SEM) framework (Millsap, 2011) has the potential to reveal contextual influences as they occur not only on the mean and variance of the latent trait or indicators, as is often explored using *t*-test, ANOVA, and MIMIC models, but also on the factor structure of classroom teaching observation instruments, which includes structural

parameters: factor means; variance and covariance; factor loadings; indicator means; and indicator variance as estimated within the overall factor structure of the observation data. The analysis of contextual variables under the SEM framework becomes valuable in aiding the development and redevelopment of measures of teaching quality. Whether conducted using multiple group modeling, multilevel modeling, or by one of the various ways the two can be combined, testing structural invariance is key to developing instruments that produce more valid results across contexts for both formative or summative purposes. This kind of rigorous analysis is of particular import when classroom teaching quality ratings are used to determine teacher pay, remediation, or termination as these kinds of high-stakes uses demand stringent validity evidence.

Establishing Evidence of Structural Validity

Establishing internal structure is one of five primary types of evidence that is relevant in building a case for validity as designated by *The Standards for Educational and Psychology Testing.* According to these standards, validity evidence based on internal structure is defined as "the degree to which the relationships among test indicators and test components conform to the construct on which the proposed test score interpretations are based" (American Educational Research Association et al., 2014, p. 16). In basic psychometric theory, the test components spoken of in the standards are behaviors "representing the underlying (presumed) construct" (Raykov & Marcoulides, 2011, p. 10). A construct is another name for a latent trait which, by definition, cannot be directly observed, but can be represented by observable indicators variables. The relationship between these indicator variables is measured in much the same way as one would measure the relationship between theoretically related observable variables. Instead of using multiple regression to estimate relationships between observed variables, the correlation

between the indicators is used to establish a relationship with a latent variable. This process, known as factor analysis, can be performed as an initial exploration of the relationship between indicators and latent traits as is done in exploratory factor analysis (EFA) and/or it can be done using confirmatory factor analysis (CFA) in the presence of either strong theoretical evidence for the factor structure or after an EFA has been performed to establish the relationships between behavioral indicators and latent traits (Raykov & Marcoulides, 2011).

In the decade after *A Nation at Risk*, even as the demand for higher standards and more standardized teacher evaluation significantly increased little attention was paid to determining the structural validity of the instruments used to measure these standards. For example, in *Psychometrics of Praxis III: Classroom Performance Assessments,* which outlines the extensive multi-year development of a classroom teaching observation instrument, there is no mention of factor analysis or invariance testing. Of the 600 classroom teaching observation studies conducted in that timeframe, only four include factor analysis as a part of the study. This does not mean that factor analyses were not conducted during this time period; it does mean that they were not often formally reported. The purpose in highlighting the scarcity of factor analysis in the classroom teaching observation literature in the 1980s and early 1990s is not to disparage those who created the instruments, nor those who researched them; rather, it is to establish the context surrounding the development of the instrument used in this study.

The JPAS was developed initially as a means by which to evaluate teachers within the District so that decisions about employment were based on empirical evidence as opposed to principal perception. Prior to this, observations were significantly more subjective. For this reason, great care was taken to establish a committee composed of researchers from the University of Utah, experts in the field of teaching and learning at both the District and at the

Utah State Office of Education, and psychometricians. It was developed using an iterative process wherein a framework was first established that included theories on the way classrooms should be managed, theories on delivering instruction, and theories on the way in which teachers and students should interact. The JPAS, if analyzed closely, has many similarities to the FfT, which is not surprising given that it was developed during the same time period and is likely based on similar instructional theories.

One important feature of the JPAS classroom instructional observation instrument was that it was not a stand-alone set of principal observations, but a component of a framework that included yearly trainings for teachers on what to expect and prepare for, yearly trainings for principals on how to effectively use the instrument, the observations themselves, and an interview process in which principals gave feedback to teachers on the different domains, taking not only from the scores, but also from the notes they had made during the observation. Also included in this process was a portfolio element in which teachers provided evidence to principals of their lesson plans, assessments, assignments, professional development, and communication with parents and students. It was used summatively for all teachers, and formatively for teachers who had been in the district for less than three years. In the case of summative use, JPAS classroom teaching ratings and the notes that principals created during the observation time frame were inconsistently followed up by professional development. The onus of improvement was put on the teacher. When the JPAS classroom instructional ratings were used formatively, newer teachers were more frequently provided with professional development and mentoring that focused on improvements in areas where teachers received poor ratings.

During the time that the JPAS was developed, little emphasis was placed on the use of empirical analysis as a component of the developmental process when creating classroom

teaching observation instruments to measure this quality. In acknowledging flaws in both the instrument used in this study and the processes used to develop it, there exists also an understanding that these flaws were common in the field during this time period and that the JPAS, in many ways, met or exceeded the developmental rigor of other classroom teaching observation systems at the time.

With that stated, it was still important to acknowledge the concerns outlined in the next section. Examining these concerns reinforced the need to establish the current factor structure of the JPAS at both the classroom and the rater level. In addition, systematic examination of previous assumptions about the uniformity of classroom instruction across grade-levels and contexts provided the exigence to test for invariance across grade-level groups. Information gleaned from this study was essential to inform future decisions regarding the behavioral indicators chosen to represent classroom instructional quality at different grade levels, how those indicators should be rated, and how the individual indicator ratings might be combined to provide factor-level ratings that are both informative and actionable. In addition, this study revealed modifications that may need to be made to the observation instrument used in the JPAS in order to strengthen the validity argument for its use as the primary component of teacher evaluation within the Jordan School District.

The Jordan Performance Appraisal System: Development and Concerns

Like the Praxis III, the JPAS underwent a rigorous multi-year development process that blended substantial contemporary research in classroom instructional quality (Capie et al., 1980; Cooley & Leinhardt, 1980; Doyle, 1986; Evertson et al., 1980; Kallison, 1986; Rosenshine, 1983; Weinstein, 1979) with the expertise of school, district and state-level educators. In addition, documentation connected to the JPAS indicates that after the system was piloted for

one year, psychometricians from the Institute of Behavioral Research in Creativity (IBRIC) performed a series of analyses on the resulting observational data (Jordan School District, 1996). In addition to establishing the statistical properties of indicator-level data such as mean, variance, standard deviation, and reliability, IBRIC also performed a principal components analyses (PCA) using SPSS (Jordan School District, 1996). While this development process was thorough and rigorous for a classroom teaching observation instrument developed during the 1990s, several concerns were substantial enough to merit a new study on the factor structure of the JPAS in the 2010s.

The first concern was that bias was created by transforming count data into categorical data from indicators whose possible counts ranged from 9 on some indicators to 60 on other indicators. Tallied data from 13 count indicators were transformed to three-category responses so that teacher ratings for count indicators could more closely resembled ratings from categorical indicators. Transforming count data in this manner removed a significant amount of variance without a theoretical justification for its removal. While this transformation made estimating factor structure much simpler, doing so without both a theoretical basis and a statistically supported algorithm comes at a cost that manifests in biased estimates and incorrect assumptions about the relationship between behavioral indicators and latent traits.

To add to the list of issues, indicators from the original instrument developed in 1996 were modified in 2013 so that the JPAS fit state requirements in educator evaluation. The newest iteration of the JPAS is based on instructional theories from the 1980s and early 1990s mixed with theories from the 2000s and 2010s. It still contains some of the more prescriptive indicators related to managing routines and listing objectives while also including indicators that relate to student behaviors, student interactions, and differentiated instruction (Appendices A-C). Some

indicators were modified to align with theories that behavioral indicators chosen to measure classroom teaching quality should include both teacher and student behaviors interlacing teacher knowledge, practices and beliefs with student knowledge, practices, and beliefs (Bell et al., 2012). After these changes, there was no indication that another analysis was conducted beyond the examination of basic statistical properties such as mean, variance, and standard deviation.

The final concern relates to the "widget effect" (Weisberg et al., 2009). A number of prescriptive indicators on the JPAS are easy to measure, but also have a high (95% or higher) rate of success. As a result, the distribution of the response data from these indicators have very little variance. Indicators with such a high level of success are problematic for a several reasons. First, they do not provide enough variance to be effectively incorporated into the measurement model: Covariance between indicators is difficult to establish when individually they do not vary significantly from the mean. This is compounded in count data in that zero-inflation makes linking count outcomes difficult, require an extra parameter in an already complex measurement model. Second—and more important to district personnel who rely on results to make decisions about remediation and professional development—other than identifying a handful of the least effective teachers, these indicators do not give sufficient information to aid in efforts to improve classroom instructional quality through professional development and mentoring. When most are getting an exceptional or perfect rating, information on what should be improved is sparse. This reflects one of the biggest flaws in current observation systems: "the precedent of not differentiating among teachers" (Cohen & Goldhaber, 2016, p. 1). A lack of indicator-level variance is problematic not only because of its impact on an analyses—low-variance indicators can obfuscate the relationships between other indicators—but because every indicator takes both time and attention to rate. Indicators that do not provide substantial information about a teacher's

classroom instructional quality take time away from other indicators that are more sensitive to the latent traits of interest in the instrument. If principals have only a limited number of behaviors that they can observe in a given time-frame, then each of the sample behaviors should be difficult enough to differentiate between a teacher with low, moderate, and high levels of the trait of interest.

It is important to note that since the original development of the JPAS, many methodological advances have made it possible to better analyze count data. These advances combined with increasing rigor in the study of the factor structure of different measures of teaching have opened up not only the possibility, but also the demand for this study.

Research Purpose

This study examined the factor structure of the classroom teaching observation instrument used in the JPAS. This study served both a functional and a theoretical purpose. It provided information to District personnel who will use it to make decisions regarding future research, development, and uses of the observation instrument. It added to a growing body of research on (a) locally developed instruments; (b) multilevel factor analysis of classroom teaching observation data as an important component of a validity argument; (c) invariance testing within the SEM framework; and (d) estimating factor models with Poisson and negative-binomial distributions by answering three specific research questions.

Research Questions

1. What factor structure best represents the underlying relationship between the JPAS behavioral indicators of classroom instructional quality when used in Grades 1-6?
    a. To what degree does the model indicate a unidimensional construct of classroom instructional quality?

        i. What percentage of the variance within each indicator is explained by the latent variables they represent?

    b. If the structure is not unidimensional, how many factors are represented by the behavioral indicators?

        i. To what degree do the latent variables correlate with one another?

2. What percentage of the variance in the latent variables of the model is explained by the variability between raters?
3. To what degree are the results of a confirmatory factor analysis performed on behavioral indicators invariant across grade groups?

CHAPTER 2

Review of the Literature

This study was conducted as a component of what Sirotnik (1980) refers to as the psychometric phase of research: Establishing the psychometric qualities of an instrument. This study examined the factor structure of the classroom teaching observation instrument used in the JPAS. In order to provide a relevant and focused review of the literature, articles and reports were examined for studies related to the measurement of observed classroom instructional quality that specifically examined the factor structure of classroom instruction observation instruments. Studies that were stand-alone or in conjunction with studies where results were used to examine something other than the instrument itself were both included. As the purpose of this review was to catalogue the increase in sophistication of factor analysis as it appears in the classroom instruction observation literature. The purpose of this literature review was to establish this study as a necessary element within the existing literature that continues the current trajectory of increased sophistication in methods used to analyze of the factor structure of classroom instruction observation instruments.

Literature Search Procedures

The literature review was conducted using ERIC, PsychInfo, EconLit, & Education Full Text. Using the thesaurus, the following search term was found to be relevant to the measurement of classroom instructional quality: *classroom teaching observation techniques.* A search of this terms within published articles and dissertations between 1980 and 2020 yielded 6,283 results. A second search was conducted adding in the thesaurus terms *factor analysis*, *factor structure*, and *psychometrics*. Combining the two searches yielded 73 results which were further filtered to include only academic journals and dissertations. Those 73 articles were then

screened for relevance. Relevance was determined based on the inclusion of *factor analysis, factor structure, mulitlevel factor analysis, or invariance testing.*

Results

Of the 73 original articles and reports, 32 were found to be relevant enough to include in the literature review. Once these articles were deemed relevant enough for use, the reference pages from each of the articles were used to find studies that may have been missed in the initial search, and additional studies were added to the original 32. The studies from relevant searches are presented by analysis type in order to illustrate the manner in which methodologies progress, and also to allow for discussion of strengths and weaknesses evident in the literature.

Principal Components Analysis

During the two decades after *A Nation at Risk*, researchers rarely looked beyond initial theoretical approaches to classroom teaching observation systems in order to analyze the structure of the instruments being used to measure classroom teaching observation data. Those that did more often than not used principal components analysis (PCA) as the method of extraction (Beem & Brugman, 1985; Crocker & Brooker, 1986; Jordan School District, 1996; Pilburn & Sawada, 2000). In research where latent variables are not correlated, PCA may produce similar results to factor analysis if (a) the communalities are close to 1.0 and (b) there is a large number of variables. (Bandalos, 2018) PCA "transforms an original set of variables into a substantially smaller set of uncorrelated variables" (Dunteman, 1989, p. 7). This type of analysis does not indicate the degree to which a factor contributes to an indicator rating because PCA assumes that the communality is 1.0 or very close to it. If one chooses to use PCA, it needs to be done with the assumptions of the analysis in mind.

In developing the Reformed Teaching Observation Protocol (RTOP; Pilburn & Sawada, 2000) researchers utilized PCA as the extraction method. Observational ratings from 25 indicators collected from 153 classrooms were analyzed in SPSS using PCA and a Varimax rotation. Varimax, which is an orthogonal rotation, assumes that the correlation between latent traits is zero. After conducting the analysis, researchers noted that while three latent traits appear to be indicated, "many indicators are not uniquely identified with a single factor" (Pilburn & Sawada, 2000, p. 20). This highlights the importance of understanding assumptions before using specific rotations. The shared variance discovered by researchers in the development of the RTOP is indicative of the need for oblique rotation yet an orthogonal rotation that was used. Similar issues exist within the JPAS analysis reported in the 1996 JPAS development literature. In addition, it is highly likely given the nature of classroom observation data and the impact of rater that the communalities were not actually 1.0 and that some important residual variance existed that was unaccounted for.

While less frequent than in decades prior, PCA has still been used within the last 10 years, particularly during the timeframe between NCLB and the ESSA when research into the psychometric properties of classroom teaching observation instruments took a secondary position to research questions related to student growth models such as VAMs and SGPs. To examine the structure of the PLATO (Grossman et al., 2013) in order to determine the degree to which ratings from the PLATO could predict VAM results for the same teachers, researchers conducted what they believe to be an EFA using principal component analysis as the extraction method when in fact, they had used PCA. It is important to note that PCA is not an extraction method of EFA as mentioned in the study, but a separate type of analysis that relies on separate assumptions (Raykov & Marcoulides, 2011). As with other studies, the PLATO study originally used

orthogonal rotation. This analysis was followed up later with other studies wherein EFA was utilized and the proper rotation employed (Grossman et al., 2014). These follow-up studies used a more appropriate process for classroom observation data as described in the next section. The main reason why EFA is more appropriate generally speaking when analyzing data from classroom teaching observations is that rater effect has been indicated as a significant source of variance, making communality unlikely to be close to 1.0.

Exploratory Factor Analysis

Exploratory factor analysis (EFA) has existed as a methodology for over almost a century (Spearman, 1904, 1927). Since that time, multiple studies have verified that EFA is an effective tool in establishing the factor structure where strong theoretical evidence for a structure does not exist or has come into question (e.g., Fabrigar et al., 1999; Ford et al., 1986; Gorsuch, 1990; McNemar, 1951).

Two studies that were conducted more than a decade after the publication of *A Nation at Risk* indicate that EFA was employed as the reduction method for the study of the factor structure of two separate observation instruments (Chauvin et al., 1991; Manaf, 1995). As is often the case, the difference between methodologies is a choice, whether conscious or not, to make assumptions about the nature of different parameters (Gorsuch, 1990). The decision of which rotation to use, along with other decisions regarding EFA—which variables to include and how many latent traits to retain—are important to producing valid and reliable results when analyzing factor structure (Fabrigar et al., 1999). Unfortunately, as Fabrigar et al. point out, "researchers appear to be unaware of the issues involved in these decisions" (1999, p. 273). This lack of awareness appears in some of the classroom teaching observation literature in the form of small missteps that can bias results.

Alongside these studies that made missteps, there exists many studies that indicate an awareness of the issues involved in the decisions made during factor analysis, a study on the factor structure of the CLASS and the MQI conducted using a population of 390 fourth- and fifth-grade students and their teachers included EFA as a method to establish the factor structure of both instruments and followed a careful path of decision making while conducting the EFA (Holmes & Bolin, 2017). Even nearly three decades prior to this study, researchers from Louisiana State University conducted an EFA with oblique Promax rotation using the SAS program to analyze the *System for Teaching and learning Assessment and Review* (STAR; Ellett et al., 1991). By employing oblique rotation, using careful analysis of eigenvalues and loading patterns, researchers combined robust empirical knowledge with theoretical knowledge to estimate the factor structure of the STAR (Ellett et al., 1991). In doing so they provided an example of the way in which data from classroom teaching observations can be carefully and thoughtfully analyzed. In addition to taking careful steps throughout the EFA process, researchers also cross-validated results by conducting a CFA using a new sample of data from the same population, a step often missed in classroom teaching observation studies of factor structure.

Many studies conducted after initial analyses have caught some of the problems of earlier studies. In a series of analyses of the Observer Rating Scale (ORS; Briggs & Dickersheid, 1985), researchers analyzed data from classroom teaching observations a decade after the original development in order to explore the instrument's purported factor structure. The original ORS included nine indicators of teacher personality and behavior: (a) enthusiasm, (b) warmth, (c) feedback, (d) on-task activity, (e) cognitive demand, (f) variety, (g) freedom, (h) individualization, and (i) clarity. By employing an oblique method of rotation, in the case of this

study, Promax, the analysis took into account the correlation between factors while the extraction of the factors was done using unweighted least squares. Findings from the analysis revealed a four-factor structure, which was significantly different from the purported structure that the instruments' developers, using theoretical information alone, claimed represented the data (Manaf, 1995).

The importance of this study is its contribution to the understanding that what researchers and practitioners may conceptually theorize to be the relationship between indicators and latent traits may not fit the empirical relationship established through EFA. Theoretical relationships developed by content experts are essential to the process of development and should not be dismissed based on the results of psychometric analyses, but the use of psychometric analyses is key in providing evidence to help build stronger, more defensible theoretical structures rather than relying solely on a priori evidence (Fabrigar et al., 1999; Rakov & Marcouledes, 2011). The two are reliant on one another. Whether established through theory or through EFA, it is essential that a follow-up analysis be conducted (in the case of EFA, using a separate data from the same population) in order to determine whether the results are can be cross-validated.

Confirmatory Factor Analysis

In determining whether or not to perform a CFA, it is important to establish that CFA has a specific purpose that is related to but should not be interpreted as the purpose of the EFA (Brown, 2015). As noted in an Mplus discussion on CFA:

> CFA is appropriate in situations where the dimensionality of a set of variables for a given population is already known because of previous research. The task is not to determine the dimensionality of a set of variables or to find the pattern of the factor loadings. Instead, CFA may be used to investigate whether the established dimensionality and

factor-loading pattern fits a new sample from the same population. (Muthén & Muthén, 2020).

In some earlier studies, researchers attempted a CFA to determine if results could be cross-validated but did not use a separate sample from the same population, ergo they merely analyzed the same data using a different method (Manaf, 1995). Some studies followed the proper procedure by using a new data set to determine the degree to which the factor-loading pattern from the original analysis fit a different sample from the same population (Holmes & Bolin, 2017; Manaf, 1995). While this could be accomplished using a second EFA on a separate data set, the benefit of the CFA is that it allows loading parameters to be fixed at specific values where theoretical or empirical evidence indicates that the relationship between an indicator and a factor is weak. It also allows for factorial invariance testing where the ability to fix factor loadings to be equal, or to hold a specific value is essential to model comparisons.

Many studies conducted in the 2010s focused on repeated analysis of instruments in different contexts than those of the original instrument's development. For example, the factor structure of the CLASS (Pianta et al., 2008), which now exists in multiple forms for toddlers, pre-K, lower, and upper elementary as well as secondary—has been analyzed in various pre-K (Mashburn et al., 2006) and elementary populations in the United States (Sandilos et al., 2016) as well as in secondary settings (Pianta et al., 2008; Malmberg et al., 2010; Hafen et al., 2015; Lockwood & McCaffrey, 2009) and with English language learners (Downer et al., 2012). Internationally, the factor structure of the CLASS has been reexamined in populations of students and teachers in different countries such as Finland, China and Norway (Hu et al., 2016, Pakarinen et al., 2010; Virtanen et al., 2018; Havik & Westergård, 2020).

One study indicated that the structure of an instrument maintained similar relationships between behavioral indicators across different populations (Hu et al., 2016), other studies, such as the CFA conducted using data from 417 kindergarten classrooms, indicated a very different structure from that presented in reports and handbooks for the instrument (Sandilos & DiPerna, 2014).

Two studies focus on the factor structure of multiple instruments in order to make comparisons from the results. For example, the Early Childhood Environment Rating Scale-Third Edition (ECERS-3) used the CLASS as a comparison. This study was conducted across three states using a large sample of classrooms where only data from the ECERS was collected from the large sample while data from the CLASS Pre-K and the ECERS-3 were both collected from a subset of 119 of those classrooms in order to study the relationship between the two instruments. These kinds of comparisons help to build upon structural validity evidence in order to establish criterion validity and are particularly important when developing newer classroom observation instruments or when making an argument to use one over the other (Virtanen et al., 2018).

At least two studies used CFA to examine the possibility of a bifactor model (Crawford et al., 2013; Sloat et al., 2017) wherein each indicator loaded on a general factor in addition to multiple sub-factors. These analyses are especially important in that they exemplify the manner in which a common trait of instructional quality can be measured simultaneously alongside multiple traits.

Multilevel Exploratory Factor Analysis and Confirmatory Factory Analysis

While the studies mentioned previously examine observation data at the classroom level without taking into account the impact of the school or rater within the model, McCaffrey et al.

(2015) make a compelling argument for the use of multilevel models. By clustering teacher ratings by rater, researchers were able to establish the structure of the CLASS while providing evidence of significant rater-level variance.

Invariance Testing in a Multilevel Model

As there are no studies in the classroom instructional quality literature that examine within-level group variance in multilevel models, examples from studies outside of this literature were used to guide the process (Asparouhov & Muthén, 2012; Kim et al., 2012; Kim et al., 2015; Ryu, 2015). From these studies, the basic principles of examining within-level factorial invariance were established.

These studies each proposed different steps with this process with different foci based on the research questions being asked, and with the understanding that there is no single approach to determining measurement invariance at the within level of a multilevel model. One study focused on using a multiple indicator multiple cause (MIMIC) model where within-level variables served as observed predictors of the latent variable and were treated as covariates in the structural model. The weakness of this model is its inability to address the possibility that the cluster variable may impact the overall structural parameters within each group such as loadings or indicator-level means differently (Kim et al., 2015). A second study treated a group as an exogenous variable (Jak et al., 2014), and third study proposed "a multigroup MSEM framework (called MG1-MSEM) that uses Muthén's limited information maximum likelihood (MUML) estimation" (Ryu, 2015). This approach is sensitive to cluster size, and estimates can be affected when cluster sizes are not balanced, making it a poor approach for this study, which does not have data with balanced cluster sizes. Also, this approach does not allow for school-level random effects which should be taken into account as a possible source of bias, specifically in this study

because raters, all of whom have taught at different levels over the course of their pre-administrative careers, may be more lenient or more severe depending on the grade level of the teacher being evaluated. For these reasons, the approach was not used.

One study also explored the option of an nSEM framework using R package xxM (Ryu & Mehta, 2017). While the nSEM has benefits in cases of "complex data structures that could introduce additional complexities in the standard MSEM framework, such as cross-classified data, partially nested data, and longitudinal data with switching classification, (Ryu & Mehta, 2017, p. 938), none of these complexities were an issue in the data used for this study, and no known studies indicate which program or method is best for Poisson and negative binomial models.

Summary

Classroom teaching observation literature over the last 30 years reveals a trend of increasing rigor in studies that involve establishing structural validity. Studies in the 1980s and 1990s are overwhelmingly conducted using PCA, but beginning in the 1990s, and more prevalently in the 2000s, models show increasing sophistication as EFA and CFA have become the dominant methods. In the 2010s MCFA began to emerge as a means to control for rater and school-level bias. With each stage, an increasing amount of residual variance has become estimable. While some research falls back to prior mistakes, the general trend seems to move into increasingly complex modeling techniques with greater attention to the importance of each step along the way. Studies on invariance within the multilevel SEM framework were missing from the literature, highlighting the importance of this study in exploring important questions related to measuring classroom instructional quality amidst student, teacher, and classroom-level variables.

CHAPTER 3

Method

This study analyzed the factor structure of the classroom teaching observation instrument used in the JPAS using EFA, CFA, MCFA, and MLMG-CFA in progressively more complex models meant to answer all three of the research questions.

1. What factor structure best represents the underlying relationship between the JPAS behavioral indicators of classroom instructional quality when used in Grades 1-6?
    a. To what degree does the model indicate a unidimensional construct of classroom instructional quality?
        i. What percentage of the variance within each indicator is explained by the latent variables they represent?
    b. If the structure is not unidimensional, how many factors are represented by the behavioral indicators?
        i. To what degree do the latent variables correlate with one another?
2. What percentage of the variance in the latent variables of the model is explained by the variability between raters?
3. To what degree are the results of a confirmatory factor analysis performed on behavioral indicators invariant across grade groups?

The following section describes in detail the study's methodological and procedural elements. It includes the study's design, participants, instrument, data collection, and analysis.

Design

This study used a multi-year cross-sectional design: Though data were gathered over a three-year time period, and each teacher was observed at different points during that time period,

only a single set of observations at a specific point in time was included for each teacher. In the Jordan School District, principals observed teachers giving instruction to their students during a 30 to 45 minute segment of class time in a two-occasion set of observations—the second occasion occurring within two weeks of the first. Most teachers were observed once every three years. Provisional teachers who had taught less than three years in the Jordan School District were observed more often, but only the most recent set of observations was included in this study. The most recent observation set was the one used in the teacher's final rating. Using more than one set of observations for a teacher could bias results by over representing a specific rater or teacher.

Participants

At the time of the study, the Jordan School District student population was composed of 52,600 students of which 21,500 were elementary students in classrooms of teachers participating in this study with a teacher to pupil ratio of 1 to 24. Of this population, 22% were on free or reduced lunch, 10.8% received special education IEP accommodations, 5% were classified as English language learners, 2.5% were classified as homeless, and less than 1% were classified as immigrant or migrant. In addition, 2.8% were American Indian or Alaskan Native, 3.6% were Asian, 2.7% were Black, 3% were Pacific Islander, 91% were White, and 14.4% were Hispanic ethnicity. The gender of students was nearly evenly divided with 50.8% female students and 49.2% male students. While data exists on all grade levels, elementary and secondary classrooms are structured very differently. Most elementary teachers have the same students all day long and teach multiple subjects. Most secondary teachers have students for what amounts to 45 minutes per day with high school levels teaching for 90 minutes every other day. They teach each group of students in a specific content area and each group of students is

unique from the other. In addition, the grade-level groupings are already assigned to different sets of principals because of the division between middle and high schools, making the methodology for exploring the questions posed in this study different. At the secondary level, the group-level variance occurs only at the between level, not the within level. Elementary was chosen first because district personnel wish to make instrument changes at the elementary grade levels first before moving on to the secondary grade levels.

Participants included 51 elementary school administrators. Of the 34 elementary schools these principals worked in, 7 were classified as Title 1 schools. In addition, 10 of the administrators were male, 31 were female.

Participants also included all teachers of Grades 1-6 who were eligible for educator evaluation. This included licensed part-time and full-time contracted teachers. It did not include interns, student teachers, or teachers who worked hourly. The study spanned three school years: 2014-2015, 2015-2016, and 2016-2017 enabling the inclusion of all 1220 non-hourly elementary teachers. Of the 1220 teachers used in this study, 345 teachers were employed at the district for three years or less and were considered provisionary. In addition, 150 were male and 1170 were female. Approximately 620 taught Grades 1-3, and 600 taught Grades 4-6. About 200 teachers were employed as full-time special education teachers.

Instrument

The JPAS classroom teaching observation rating instrument was used in this study. The instrument was initially developed in 1996 by the Jordan School District in consultation with the IBRIC but was revisited in 2013 in order to ensure compliance with state educator evaluation requirements. The 2013 classroom teaching observation rating instrument used in this study was composed of 49 indicators that were intended to measure three constructs: The 13 indicators in

Domain I purported to measure classroom management; the 25 indicators in Domain II purported to measure the delivery of instruction; and the 11 indicators in Domain III purported to measure the teacher's interaction with students (See Appendices A-C).

Observation

The data from the classroom teaching observation section of the JPAS represented two observation occasions. Teachers were given two weeks advanced notice that they would be observed by an administrator from their school. After the two weeks, the administrator came to their classroom without giving any further notice to the teacher. Teachers who felt unprepared could ask the principal to come back another time. Teachers who were conducting activities that did not include teaching at that time could also ask the principal to come back another time. This opportunity to postpone was allowed only once per teacher. The first and second occasions in an observation set occurred within two weeks of each other. For teachers who were evaluated once every three years, the data used in this study represent the two-occasion observation set from the most recent year that a teacher was evaluated. For teachers who had been in the district less than three years, it represented the final two-occasion observation set of that school year: Any other observations performed during the year were excluded from the study data.

Upon arriving in the classroom to observe the teacher, the rater recorded the start time and the number of students in the classroom before beginning the observation. During the observation, the rater tallied or assigned a rating for each behavioral indicator of classroom instructional quality. Principals used a rubric designed by the Jordan Education Committee to guide them in this process. Count indicators were tallied as behaviors were observed. Raters also took notes during this process. Notes often included drawings of the classroom layout or remarks when a teacher performed a behavior well or failed to perform a behavior well. These notes, in

addition to tallies and other ratings, were used at the end of the observation in order to fill out ratings for summary indicators.

Data Collection

Jordan Evaluation System (JES) personnel scanned the forms into a machine that was connected to a computer housed in the JES office. This computer included a program which transferred the data to a database housed in the Jordan School District Main Office. The data were scanned as forms were submitted to the JES. Forms were also checked manually to ensure that circles were filled in completely and were readable by the scanning machine. In addition, forms were reviewed in order to ensure that the correct information had been filled in for each teacher. In the event that indicators had been left blank, the rater was asked to fill in the appropriate information based on notes taken and tally marks made so that every educator had a complete evaluation. In spite of these precautions, a few pieces of data were missing.

Data Preparation

Prior to analysis, data were divided into two subsets. In order to facilitate stratified random selection and ensure that all grade levels and schools were equally represented in each data set, data were stratified by grade and school. Once data were stratified, they were assigned randomly to one of two groups. The purpose of the two separate data sets was to have one data set for an EFA and one data set for a CFA. Each sample contained over 1000 observations. The number of observations was sufficient to perform EFA and CFA analyses at both the between and within level (Gagné & Hancock, 2006). In order to prepare data for analysis using Version 8 of the Mplus program, all data were converted to numerical form. Names and text identifiers were replaced with representative numbers. Missing data, including instances where there was

insufficient opportunity for the behavior to be observed, were identified using an appropriate numerical representation suitable for Mplus: 999.

Distributions of ratings for each indicator were analyzed to ascertain the degree to which data for each indicator had enough variance to be considered valuable to the analysis. Data from many of the count indicators were zero-inflated. Those zero-inflated indicators where 95% or more of the teachers receive zero tally marks were excluded from any factor analysis. In addition categorical indicators where 95% or more of the teachers observed received the best rating possible were also excluded (See Table 1). Omitting indicators with low variance kept them from affecting parameter estimates for the latent variables and their behavioral indicators. Of the 49 indicators, three count and eight categorical indicators were eliminated due to insufficient variance.

Table 1

Percent of Teachers with Highest Rating or Zero Count per Indicator

Indicator	Type	%with Highest Rating	% With Zero Count
1	Count	N/A	82
2	Count	N/A	96
3	Count	N/A	99
4	Count	N/A	97
5	Categorical	99	N/A
6	Categorical	98	N/A
7	Categorical	100	N/A
8	Categorical	99	N/A
9	Categorical	90	N/A
10	Categorical	91	N/A
11	Categorical	92	N/A
12	Categorical	86	N/A
13	Count	N/A	97
14	Count	N/A	20
15	Count	N/A	11
16	Count	N/A	19

Table continues on next page

Table 1 (Continued)

Indicator	Type	%with Highest Rating	% With Zero Count
18	Count	N/A	13
19	Count	N/A	18
20	Categorical	62	N/A
21	Categorical	83	N/A
22	Categorical	99	N/A
23	Categorical	99	N/A
25	Categorical	84	N/A
26	Categorical	86	N/A
27	Count	N/A	35
28	Count	N/A	39
29	Count	N/A	94
30	Categorical	54	N/A
31	Categorical	86	N/A
32	Categorical	95	N/A
33	Categorical	79	N/A
34	Categorical	89	N/A
35	Categorical	66	N/A
36	Categorical	88	N/A
37	Categorical	91	N/A
38	Categorical	93	N/A
39	Count	N/A	20
40	Count	N/A	30
41	Count	N/A	31
42	Count	N/A	51
43	Count	N/A	31
44	Count	N/A	68
45	Count	N/A	59
46	Categorical	70	N/A
47	Categorical	58	N/A
48	Categorical	97	N/A
49	Categorical	86	N/A

A qualitative analysis of each behavioral indicator was conducted in consultation with multiple Jordan School District employees including the administrator over Teaching and Learning at the Jordan School District, elementary principals, and members of the Evaluation, Research, and Accountability department. Indicators listed in Appendix D were not retained for (a) lacked sufficient variance, (b) lesson-dependence or (c) consisting of multiple indicators.

Analysis of Remaining Indicators

Analyses of the indicators that were retained (see Table 2) were performed in Mplus using maximum likelihood with robust standard errors (MLR) as the estimator. As noted in the Mplus User's Guide 8, "The default estimator for this [analyzing count data] is maximum likelihood with robust standard errors using a numerical integration algorithm" (Muthén & Muthén, 2017, p. 48). In the MCFA and MLMG-CFA analyses, maximum likelihood using first-order derivatives (MLF) was used instead of MLR due to the complexity of the models. MLF is equivalent to MLR with large samples (Muthén & Muthén, 2010). Our sample was large, justifying the use of MLF for the purposes of the analyses. Missing data was managed using Full Information Maximum Likelihood (FIML) which Enders (2010) indicates to be a robust method of managing data that is not missing completely at random. In the input for the analysis, FIML was indicated by using the term MISSING = ALL (999).

Table 2

JPAS Indicators Retained

Indicator	Indicator Description	Indicator Type
9	Low-key tactics for misbehavior are used effectively.	Categorical
10	Teacher identifies those who are initiating the disruptions in order to end them quickly.	Categorical

Table continues on next page

Table 2 (Continued)

Indicator	Indicator Description	Indicator Type
11	Classroom routines are outlined and followed.	Categorical
14	Teacher asks factual questions to assess learning	Count
15	Teacher explains an academic concept.	Count
17	Teacher illustrates a relationship by tying new information to concepts students understand.	Count
18	Teacher emphasizes an important point in the lesson.	Count
24	Teacher displays clearly discernable interest in the subject matter through speech and body language.	Categorical
25	Teacher explicitly states goals, objectives, and expectations and relates them to the learning activity.	Categorical
26	Teacher helps to deepen student understanding.	Categorical
27	Teacher incorporates higher level thinking questions.	Count
28	Teacher asks a question and pauses for at least three seconds before calling on a student.	Count
29	Tally for each time the teacher sustains dialogue with a student by asking follow-up questions.	Count
31	Teacher uses instructional strategies that incorporate higher-order thinking skills.	Categorical
35	Teacher prepares students for activities using directions and ensuring students understand those instructions.	Categorical
38	Teacher monitors and guides all student learning to help them increase level of performance and understanding.	Categorical
39	Teacher initiates an interaction with a different student about the academic content of the class.	Count
40	Teacher provides academic feedback	Count
41	Teacher uses a procedure to get student attention before moving forward in the lesson.	Count
42	Teacher recognizes a student who is not participating and solicits their involvement.	Count
43	Tally is recorded if the teacher offers specific praise to students.	Count
44	Teacher acknowledges or praises the effort a student has made learning new material.	Count

Question 1: Factor Structure of the JPAS

In order to explore the number of latent variables the data represent, a series of EFAs were performed (Fabrigar & Wegener, 2012). The initial EFA began with the 31 behavioral indicators listed in Table 4. Each measured a different behavior, which could be expected to be observed in any given 30-minute time frame, and exhibited sufficient variance (> .95) to merit inclusion.

When determining the number of latent traits to analyze, two considerations were taken into account: First, how many latent variables were theoretically present, and second, how many latent variables could be managed in the presence of count indicators given the sample size. Sufficient evidence for three latent variables lead to a decision to test four models: Single-factor, two-factor, three-factor, and four-factor. The input TYPE = COMPLEX EFA 1 4 option in the ANALYSIS command was indicated to reflect this decision. Once the possibility of a four-factor model was eliminated, the TYPE = COMPLEX EFA 1 3 option was indicated in the analysis command using only 13 of the original indicators (See Table 3). For both EFA Model 1-4 and EFA Model 1-3, The TYPE = COMPLEX option in the ANALYSIS command with rater as the cluster variable was used in order to model the nesting of the data within raters. Multilevel EFA was not available for count data.

Models were chosen based on best fit as indicated by lower BIC values. Goodness of fit for both EFA and CFA models would normally rely in part on both absolute fit indices such as RMSEA and SRMR as well as comparative fit indices such as TLI, and CFI (Tucker & Lewis, 1973; Bentler, 1990; Hu & Bentler, 1999). The analysis of count data provides no covariance matrices and no means by which to calculate these fit indices.

A CFA was conducted after the model was established for the EFA. This was done in order to determine whether the structure established through the EFA could be cross-validated (Brown, 2015). Unlike EFA, CFA allows for factor loadings to be fixed to a specific value or freely estimated. In making model decisions, it was important to take into account automated decisions made by default in Mplus. These automated decisions included factor loadings from the latent traits being fixed to zero unless indicated in the input through the BY term connecting behavioral indicators to specific latent traits. Additionally, one of the defaults in the Mplus program indicates that the variance of each latent trait is freely estimated while the first factor loadings for each latent trait is fixed to 1.

While the mean to variance ratios of most of the count indicators would be evidence of a negative binomial distribution (see Table 4), the first CFA model (1a) was run without including the negative binomial option in the input. This was done intentionally to illustrate the way that misspecification of count distributions can impact the overall fit of the model. It also served to test the degree to which dispersion affected model fit. In Model 1a, the dispersion parameters that are fixed to zero by default were maintained and no additional input was added to indicate estimation of the dispersion parameter. Model 1b included a dispersion parameters using the (nb) input next to the variables listed under COUNT.

Dispersion parameters are valuable from a measurement standpoint because estimating dispersion reduces bias in parameter estimates such as factor loadings, latent variable intercepts and variance. Additionally, dispersion parameters were a component of the equation used to determine the intra-class correlation (ICC).

Table 3

Count Indicators: Variance to Mean Ratio

Indicator	SD^2	*M*	SD^2/M
14	0.53	2.30	0.23
15	0.46	2.37	0.19
17	6.06	2.02	3.00
18	9.44	3.04	3.11
27	9.44	2.40	3.93
28	10.10	2.65	3.81
29	13.67	4.86	2.81
39	2.45	0.48	5.15
40	130.23	21.05	6.19
41	7.41	3.04	2.44
42	2.00	1.03	1.94
43	21.41	4.73	4.53
44	2.67	0.71	3.76

Question 2: Rater-Level Variance

Indicator-level ICCs were used to determine the need for a multilevel model (Koch, 2006). In the case of a negative binomial distribution, calculating the ICC follows a different formula than data with a Gaussian distribution. Nakagawa et al. (2017) suggest that the following formula be used to calculate the ICC for negative binomial distributions:

$$\mathrm{ICC}_{\text{P-ln}} = \frac{{\sigma_\alpha}^2}{{\sigma_\alpha}^2 + \ln(1 + 1/\lambda + 1/\theta)}.$$

In the formula, σ_α^2 represents the group variance, $\boldsymbol{\lambda}$ represents the group mean, and θ represents the dispersion parameter. Note that $\boldsymbol{\lambda}$ is bolded here to differentiate it from the λ that represents factor loadings.

This formula was used to determine individual ICCs. Overall ICCs were determined using resulting within- and between-level factor variances obtained from the output of the MCFA. Following that, a multilevel multigroup CFA (MLMG-CFA) wherein between-level effects were estimated for each group simultaneously, was used to test the degree to which rater variance was invariant across groups (Asparouhov & Muthén, 2012; Kim et al., 2015; Ryu, 2015).

Question 3: Invariance Across Grade Groups

Multiple models were utilized in order to test invariance across grade groups. To begin with, EFAs were conducted using data from each grade-level group to test hypotheses discovered during the first step of the MLMG-CFA (Ryu & Mehta, 2017). Once results indicated sufficient evidence of configural invariance, a baseline model for both grade-level groups was established. The baseline model, also called the configural model is the least constrained model. With count data, the factor loadings at both the within and between levels are freely estimated as are the intercepts at the between level. Intercepts at the within level are not estimated when modeling count data, which eliminated one of the usual steps.

In order to designate the MLMG-CFA in Mplus, the command TYPE = TWOLEVEL MIXTURE was used in the ANALYSIS section of the input. In addition, the number of groups was identified as two using the command CLASSES = c(2) within the VARIABLES section of the input. The command KNOWNCLASS = c (grade = 0 1) in the same section identified the values assigned to each of the groups within the data. Comparison models were designated using

the input commands %OVERALL% for the combined and %C#1% and %C#2% for groups 1 and 2.

According to Ryu and Mehta (2017) the parameters of interest in multilevel factor invariance are $\mathbf{\Lambda}_{kj}$ and $\mathbf{\Lambda}_{kk}$ for weak invariance, $\boldsymbol{\tau}_k$ for strong invariance. The four steps outlined in multigroup CFA followed in order to establish configural, weak, strong, and strict invariance are similar but not identical to those use for MLMG-CFA. When specifying the MLMG-CFA model, the level at which loadings and indicator intercepts are tested first, whether between or within is not important. Jak et al. (2014) began testing invariance at the within level and then moved to the between level while Ryu and Mehta (2017) began at the between level first. What matters is that "no matter which level is investigated first, we recommend that an unrestricted model (i.e. $df = 0$) is specified at the other level in the first step . . . so that the statistics are not influenced by the potential misspecification in the other level" (p. 11). Another step that was added to this analysis, given that there are no comparative nor absolute goodness of fit indices when analyzing count data, was an EFA on the two groups separately to determine if the models for the combined groups represent each group once the two are separated from each other. This step is essential when using count data due to the lack of a covariance matrix which eliminates the option of using goodness of fit indices, both absolute and comparative when determining whether the configural model has sufficiently good fit to indicate that the structure of the factors and the behavioral indicators is invariant across groups.

This was done following the same steps outlined above, beginning with the Complex EFA model and moving through to the single-factor CFAs wherein the unidimensional models for each grade grouping were tested against two- and three-factor models for cross-validation purposes.

Summary

To provide evidence of the factor structure of the observation portion of the JPAS, this study used classroom instruction observation rating data for elementary teachers Grades 1-6 in a cross-sectional design that used EFA to estimate the factor structure of JPAS rating data and CFA as a means to test whether results could be cross-validated. Rater effect was determined via MCFA, and finally factorial invariance was examined using a set of EFAs and CFAs on each of the groups separately followed by a series of MLMG-CFA models to determine the degree to which the models were invariant across groups configurally and structurally.

CHAPTER 4

Results

This section presents the results of the analyses discussed in the previous chapters as they related to the three research questions. Results from Complex EFA, CFA, MCFA, and MLMG-CFA are discussed as they pertain to the research questions.

Question 1: Factor Structure of the JPAS

EFA Model Results

The initial EFA was conducted using Geomin rotation. Indicators that fit poorly or cross-loaded equally onto two factors were removed from the model one at a time from lowest loading to highest, and the model results were reexamined after each removal. Model 1a represents the simple single-factor model. Model 1b is the complex single-factor model. Model 2 represents the complex two-factor model. Model 3 represents the complex three-factor model.

The analysis compared the simple model which ignores clustering to the complex model which takes clustering into account (see Table 4). The complex model fit the data better than the simple model (ΔBIC = -10,987.10).

The analysis also compared the three complex EFA models. The two-factor model fits better than the single-factor model (ΔBIC = -3,058.03). The three-factor model displays even better fit than the two-factor model (ΔBIC = -2,762.38).

Table 4

EFA Model Fit Statistics

Model	AIC	BIC	ΔAIC	ΔBIC
1a	104,248.10	104,518.90	—	—
1b	93,396.39	93,531.80	-10,851.70	-10,987.10
2	90,396.58	90,473.77	-2,999.81	-3058.03
3	87,611.86	87,711.39	-2784.72	-2762.38

Table 5 displays the loadings for the single-factor model. In this model, not all thirteen of the indicators load significantly onto the factor. Indicator 40 loads poorly onto the factor and indicator 43 loads poorly and negatively on the factor. Five of the indicators load only moderately onto the factor with only five of the indicators loading strongly onto the factor. This does not entirely rule out a single-factor model, but it does give evidence that it might not best fit the data. A single-factor model was included in the CFA in order to determine how it fit in relationship to other models.

Table 5

Loadings for the Single-Factor Complex EFA Model

Indicator	Indicator Description	Factor Loading
14	Asks factional questions	**0.53**
40	Gives academic feedback	0.21
15	Explains academic concepts	**0.99**
27	Asks higher-order questions	**0.97**
28	Wait time after questions	**0.97**
29	Sustains interaction with students	**0.91**
39	Initiates interaction with different students	**0.98**
17	Illustrates relationships	**0.72**
18	Emphasizes important points	**0.64**
42	Encourages reluctant students	**0.52**
41	Gets student attention	**0.40**
43	Reinforces desired behavior	-0.11
44	Acknowledges learning efforts	**0.42**

Table 6 displays the loadings and cross-loadings for the three-factor model. The correlation between Factor 1 and Factor 2 was .488, which was significant at the 5% level. The correlation between Factor 1 and Factor 3 was .150 but was not significant at the 5% level. The correlation between Factor 2 and Factor 3 was .228 but was not significant at the 5% level. Indicator 17 loads onto both Factor 1 and Factor 2.

Table 6

Loadings and Cross-Loadings for the Complex EFA Three-Factor Model

			Factor Loading		
Factor	Indicator	Description	1	2	3
Factor 1					
	14	Asks factional questions	**.92**	.53	.38
	40	Gives academic feedback	**.99**	.21	.36
Factor 2					
	15	Explains academic concepts	.29	**.99**	.55
	27	Asks higher-order questions	-.10	**.97**	.41
	28	Wait time after questions	.03	**.97**	.63
	29	Sustains interaction with students	.54	**.91**	.53
	39	Initiates interaction with different students	.11	**.98**	.28
Factor 3					
	17	Illustrates relationships	.33	**.72**	**.95**
	18	Emphasizes important points	.22	.64	**.98**
	42	Encourages reluctant students	.00	.52	**.97**
	41	Gets student attention	-.28	.40	**.87**
	43	Reinforces desired behavior	.16	-.11	**.82**
	44	Acknowledges learning efforts	.31	.42	**.99**

CFA Model Results

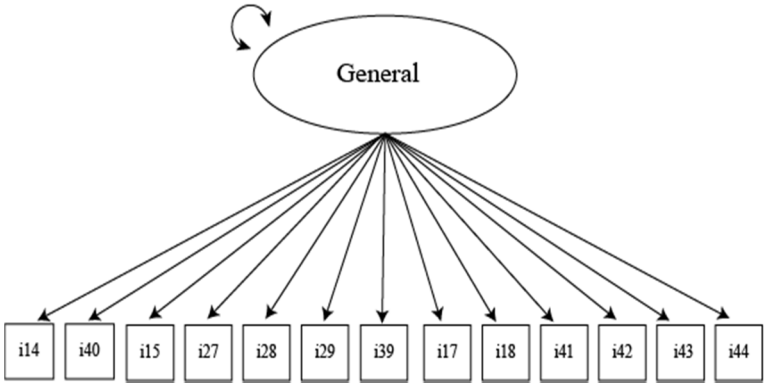
Figure 1 displays the different CFA models used in the analyses. Models 4a through 4b were CFAs with a single factor explaining all of the 13 retained indicators. Model 4a differs from Model 4b in that a dispersion parameter was not estimated in Model 4a. Instead, the model was estimated under the assumption of a Poisson distribution. While there was significant evidence that a dispersion parameter was needed, running the model without estimating the dispersion parameter allows for a comparison between models that include a dispersion parameter and those that do not. Model 4b was similar to model 4a except that a dispersion parameter was estimated to account for a negative binomial distribution.

Models 6a through 6c consisted of CFAs where indicators loaded onto three factors. The results from the previous EFA were used to determine which indicators would load onto each of the factors. As with model 4a, a dispersion parameter was not estimated for model 6a, and instead the model was estimated under the assumption of a Poisson distribution. Model 6b included a dispersion parameter just as Model 4b had. Model 6c also included a dispersion parameter, similar to model 6b. The difference between model 6b and 6c was that a constraint on indicator 17 was freed allowed it to load onto Factor 3 as well as Factor 2. This was done in order to test results from the Complex EFA and determine whether or not the cross-loadings discovered in the output from the model estimates held true once the dispersion parameter was estimated to account for the negative binomial distribution.

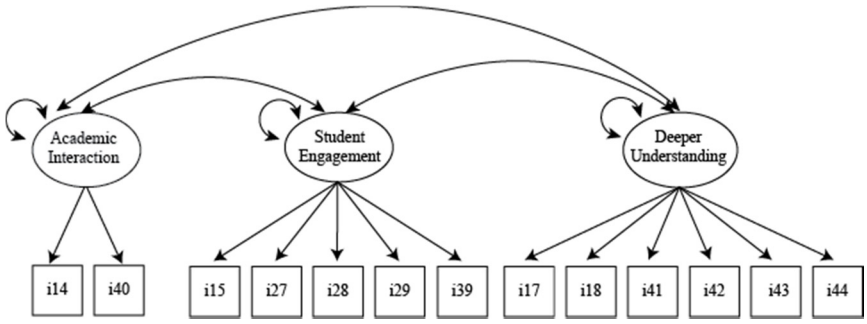
Figure 1

CFA Models 4a Through 6c

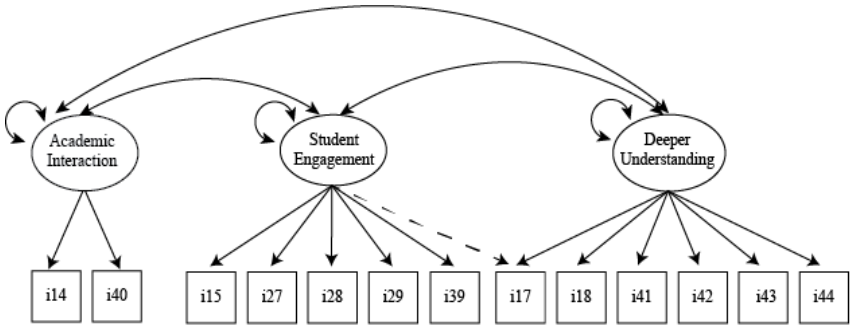
Model 4a & 4b



Model 6a & 6b



Model 6c



As displayed in Table 7, AIC and BIC fit statistics indicate that when the dispersion parameter was not estimated in the CFA model (Models 4a and 5a) the model did not fit as well as when a dispersion parameter was estimated to account for overdispersion of the data (Models 4b and 5b). Model 5c in which indicator i17 loaded onto both Factor 2 and Factor 3 did not fit better than Model 5b. As indicated by a 6.23 increase in BIC, the models are similar to one another in fit, but model 5b is the more parsimonious and the best fitting model of the two. For consistency in group-level CFA models, Model 5 would have represented the two-factor model This model lacked empirical evidence to be included in the combined-group analysis.

Table 7

CFA Model Fit Statistics

Model	LL	Parameters	AIC	BIC	ΔAIC	ΔBIC
4a	-44644.73	26	89,341.46	89,475.18	—	—
4b	-39071.49	39	78,220.98	78,421.55	-11120.50	-11053.60
6a	-42196.83	29	84,451.68	84,600.81	6230.70	6179.26
6b	-38936.36	42	77,956.73	78,172.73	-6494.95	-6428.08
6c	-38936.13	43	77,458.26	78,179.40	-498.47	6.67

MCFA Model Results

In order to determine whether or not an MCFA was needed, the ICCs for each indicator using the following formula:

$$\mathrm{ICC}_{\text{P-ln}} = \frac{\sigma_\alpha{}^2}{\sigma_\alpha{}^2 + \ln(1+1/\lambda+1/\theta)}.$$

As mentioned in Chapter 3, $\sigma_\alpha{}^2$ represents the group variance, λ represents the group mean, and θ represents the dispersion parameter.

Results displayed in Table 8 indicated that an MCFA was appropriate given the significant impact that raters had on each individual indicator. The ICC for each of the indicators appeared to be inflated. A deeper examination of group means for the 51 different clusters may

explain why the ICC was so high (See Appendix E). It could also be that equations established are not accurately estimating indicator-level ICC which has been mentioned by some researchers as a statistic that cannot always be accurately estimated (Muthén & Muthén, 2008).

Table 8

ICC Statistics: Group Mean, Variance, and Dispersion

Indicator	σ_α^2	$\boldsymbol{\lambda}$	θ	ICC
I14	10378.49	26.65	0.07	.99
I15	1042.08	2.05	0.01	.99
I17	10.85	1.98	0.54	.90
I18	107.15	3.15	0.12	.98
I27	85.60	8.46	0.94	.99
I28	922.24	26.65	0.80	.99
I29	169.67	2.05	0.04	.98
I39	51.31	1.98	0.12	.96
I40	109.85	3.15	0.12	.98
I41	465.30	8.46	0.17	.99
I42	16.86	1.98	0.35	.92
I43	17.86	3.15	0.73	.95
I44	38.17	8.46	2.10	.99

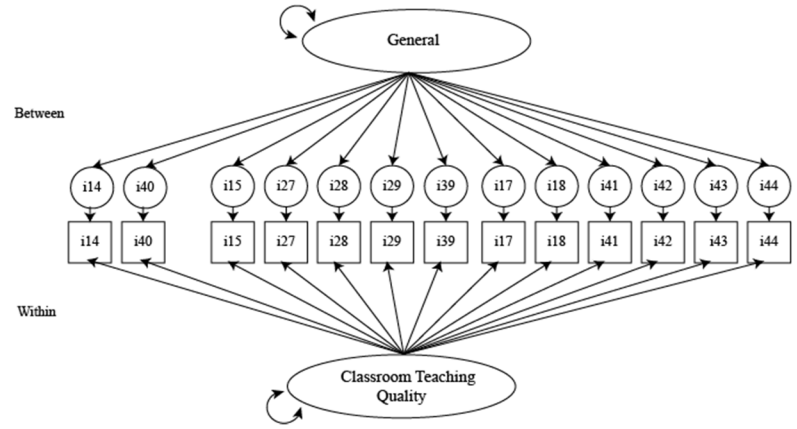
Note. The symbol $\boldsymbol{\lambda}$ is bolded to differentiate it from the symbol λ used to denote factor loadings.

As detailed in Figure 2, both the single-factor and the three-factor models were considered when conducting a MCFA to account for clustering at the rater level. This created two new models. Model 7 treated both between and within levels as a single factor while accounting for dispersion. Model 8 assumed a single between-level factor and three within-level factors while estimating dispersion. Model 9b treated both the between and the within level as a three-factor structure while estimating dispersion. Model 9a is provided as a comparison for later MLMG-CFA models as MLMG-CFA does not estimate a dispersion parameter, and so it serves as a baseline for model fit when testing measurement invariance.

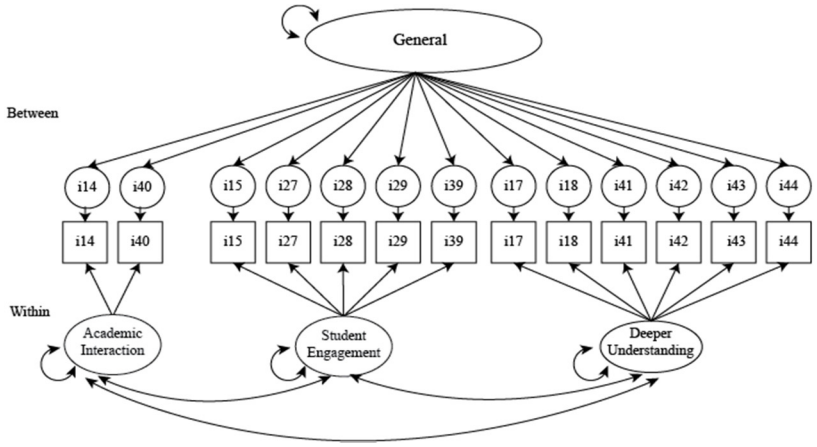
Figure 2

MCFA Models 7 Through 9a

Model 7



Model 8



Model 9 & 9a

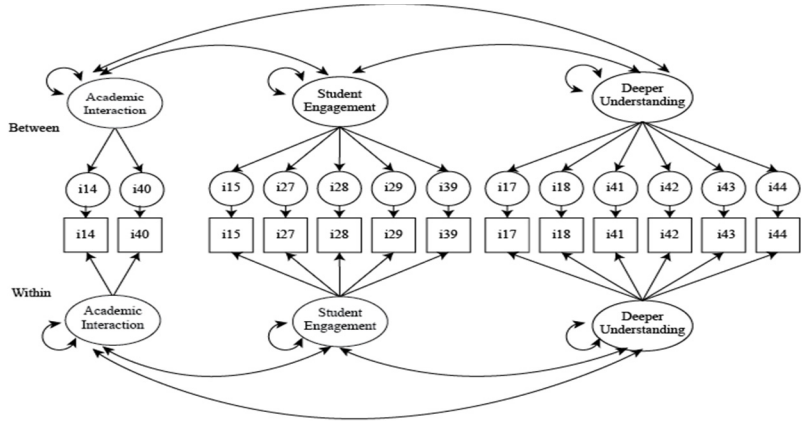


Table 9 displays the fit indices for (a) Model 6b which estimated three factors at a single level; (b) Model 7 which estimated a single-factor at both the within and between levels; (c) Model 8 which estimated a single factor at the between level and three factors at the within level; and (d) Model 9 and 9a which estimated three factors at both the within and between levels. These models illustrate the comparison between the best fitting CFA model and the MCFA models. Model 6b represented the best fitting of the CFA models. It is a three-factor model where the dispersion parameter was estimated. All MCFA models except for 5a also estimated a dispersion parameter. The MCFA models in which the distribution parameter was estimated fit substantially better than Model 6b.

As noted in Table 9, Model 9 fit better than Model 8 (ΔBIC = -719.39) and Model 7 (ΔBIC = -676.78). If it had been possible to estimate a dispersion parameter in Mplus in MLMG-CFA, Model 9 would have been the comparison model. Model 9a was retained instead for comparison purposes in later models. The inclusion of Model 9a in Table 9 illustrates the degree to which estimating the dispersion parameter changes the model fit in the presence of overdispersion. Model 9a fit was substantially worse than Model 9 (ΔBIC = 6,683.85).

Table 9

CFA Model 6b and MCFA Models 7 Through 9a Fit Statistics

Model	LL	Parameters	AIC	BIC	ΔAIC	ΔBIC
6b	-38,936.36	42	77,956.73	78,172.73	_	_
7	-38,521.56	52	77,147.12	77,414.55	-809.61	-758.18
8	-38,532.15	55	77,174.30	77,457.16	27.18	42.61
9	-38,135.24	58	76,386.49	76,737.77	-787.81	-719.39
9a	-41,196.83	45	83,261.25	83,421.62	_	_

Question 2: Rater-Level Variance

This section reports (a) within- and between-level variance, (b) the ICC for both individual factors and the overall model, and (c) results of Model 9H1—the degree to which rater effect varied between groups.

Intraclass Correlation Coefficient

The ICC used to estimate rater variance was established using the output of Model 9 (see Figure 2). The three-factor model at both the within and between levels allowed for the ICC to be estimated for each factor (See Table 10). In comparison to individual ICCs as well as ICCs for Factors 2 and 3, the ICC for Factor 1 appears to be incorrect. One of the possible reasons for the discrepancy could be that there are only two indicators that load onto Factor 1. Based on individual ICCs as well as the ICCs of the other two factors, there was empirical evidence that the ICC for Factor 1 was similar to the other factors.

Group-Specific Intraclass Correlation Coefficient

Based on the between-level variance estimated in the MLMG-CFA Model 9H1 using E1 and E2 to represent the rater effect on each model, the difference in rater effect was not statistically significant (p=.235).

Table 10

Within and Between Factor-Level Variance and ICC

Factor	Within σ^2	Between σ^2	ICC
1	0.105	0.013	.110
2	0.047	0.167	.780
3	0.035	0.255	.879
Total	0.187	0.435	.699

Question 3: Invariance Across Grade Groups

In this section, results from multiple models used to estimate measurement invariance are illustrated. They include separate group EFAs and CFAs as well as iterations of MLMG-CFA wherein parameters of interest in multilevel factor invariance: $\mathbf{\Lambda}_{kj}$ and $\mathbf{\Lambda}_{kk}$ for weak invariance, $\boldsymbol{\tau}_k$ for strong invariance (Ryu & Mehta, 2017) were freely estimated across groups in Model 8H2. Models were increasingly constrained until $\mathbf{\Lambda}_{kj}$, $\mathbf{\Lambda}_{kk}$ and $\boldsymbol{\tau}_k$ were fixed to be equal across groups.

Configural Invariance: Separate Group EFAs

This section reports the results of the separate group EFAs conducted to determine whether the factor structure was the same between the two groups. It includes fit statistics and factor loadings of (a) Model 1, the single-factor Complex EFA, (b) Model 2, the two-factor Complex EFA, and (c) Model 3, the three-factor Complex EFA. The labels G1 and G2 were added to the models to differentiate which group is represented within the discussion. Results displayed in Table 11 indicate that Model 3, the three-factor Complex EFA fit the data best.

Table 11

Comparison of Fit Indices in Complex EFA Models Groups 1 and 2

Model	LL	AIC	BIC	ΔAIC	ΔBIC
Group 1					
Model 1(G1)	-24,067.69	48,187.37	48,305.48	—	—
Model 2(G1)	-23,244.58	46,565.17	46,617.13	-1,622.20	-1,688.35
Model 3(G1)	-22,563.91	45,225.83	45,292.83	-1,339.34	-1,324.30
Group 2					
Model 1(G2)	-22,213.14	44,478.28	44,594.92	—	—
Model 2(G2)	-21,481.63	43,039.26	43,209.74	-1,438.74	-1,385.18
Model 3(G2)	-20,929.12	41,956.24	42,176.06	-863.68	-933.68

Table 12 lists the factor loadings for each indicator relative to their relationship to the three factors estimated in that model. The factors were not named as there has not been an opportunity yet to determine the theoretical basis for what these indicators might represent. Note that while all indicators load strongly onto at least one of the factors, indicators 41 and 42 load onto two different factors: Factors 2 and 3.

Cross-loadings were tested in later CFA models in order to determine whether they truly loaded onto two factors or whether these results were perhaps influenced by the dispersion of the data which was unaccounted for during the Complex EFA. As mentioned previously, this is one of the known weaknesses of EFA as it is applied to count data. Indicators were considered as loading onto two factors if the difference between the loadings was greater than 10%.

Table 12

Factor Loadings for Grades 1-3 Complex EFA Three-Factor Model

			Factor Loading		
Factor	Indicator	Description	1	2	3
Factor 1					
	14	Asks factional questions	**.75**	.48	.35
	40	Gives academic feedback	**.90**	.08	.34
Factor 2					
	15	Explains academic concepts	-.09	**.99**	.69
	27	Asks higher-order questions	-.31	**1.00**	.67
	28	Wait time after questions	-.23	**1.00**	.66
	29	Sustains interaction with students	.38	**.80**	.56
	39	Initiates interaction with different students	-.08	**.99**	.66
Factor 3					
	17	Illustrates relationships	-.06	.61	**1.00**
	18	Emphasizes important points	-.07	.46	**.97**
	41	Gets student attention	-.64	**.64**	**.83**
	42	Encourages reluctant students	-.55	**.77**	**.87**
	43	Reinforces desired behavior	.07	-.08	**.69**
	44	Acknowledges learning efforts	.16	.70	**1.00**

Table 13 lists the factor loadings and cross-loadings for each indicator relative to the three factors estimated in that model. Indicators were considered to load onto two factors if the difference between the loadings was greater than 10%. When comparing the results of the EFA for group one with the results for the EFA for Group 2, the same indicators load strongly onto the same factors. In the case of this data, however, indicator 17 possibly loads onto two factors. As with the data for Group 1, the cross-loading will be included in one of the CFA models in order to test if indicator 17 cross-loads onto two factors or whether the results from the EFA are due to other reasons such as dispersion in the data that was unaccounted for in this analysis.

Table 13

Factor Loadings for Grades 4-6 EFA Three-Factor Complex

			Factor loading		
Factor	Indicator	Description	1	2	3
Factor 1					
	14	Asks factional questions	**.92**	.53	.38
	40	Gives academic feedback	**.99**	.21	.36
Factor 2					
	15	Explains academic concepts	.29	**.99**	.55
	27	Asks higher-order questions	-.10	**.97**	.41
	28	Wait time after questions	.03	**.97**	.63
	29	Sustains interaction with students	.54	**.91**	.53
	39	Initiates interaction with different students	.11	**.98**	.28
Factor 3					
	17	Illustrates relationships	.33	**.72**	**.95**
	18	Emphasizes important points	.22	.64	**.98**
	41	Gets student attention	-.28	.40	**.87**
	42	Encourages reluctant students	.00	.52	**.97**
	43	Reinforces desired behavior	.16	-.11	**.82**
	44	Acknowledges learning efforts	.31	.42	**.99**

Configural Invariance: Separate Group CFA

CFA Results Group 1. Model 4 represents a single-factor model. Models 5 represents the two-factor model with Model 6 represents the three-factor model. In Model 6b the loading parameter was freed up on i29 so that it cross-loaded onto both Factor 2 and Factor 3. Based on the results displayed in Table 14, Model 6b had worse fit than Model 6a (ΔBIC = 83.73) and Model 6a had better fit than Model 5 (ΔBIC = -123.51), making Model 6a the best fitting model.

CFA Results Group 2. In the CFAs conducted using rating data from Group 2 (Grades 4-6) dispersion parameters were estimated by default in all models. Model 1 represented the single-factor model. Models 2 represented the two-factor model. Model 3a represented the three-factor model with no cross-loadings. Model 3b represented the three-factor model except with indicators 41 and 42 freed to load on both Factor 2 and Factor 3. Based on the results, Both Model 1 and Model 3a had better fit than Model 2. Model 3a fit the data better than Model 3b. Overall, Model 3a had the best fit of all models (See Table 15).

Table 14

CFA Model Fit Statistics Grades 1-3

Model	LL	Parameters	AIC	BIC	ΔAIC	ΔBIC
4(G1)	-20,256.94	39	40,591.89	40,766.79	_	_
5(G1)	-20,251.62	40	40,538.24	40,762.62	-53.65	-4.17
6a(G1)	-20,183.38	42	40,450.76	40,639.11	-87.48	-123.51
6b(G1)	-20,246.51	45	40,642.32	40,721.84	171.56	82.73

Note. Dispersion parameters were estimated for all models in this table.

Table 15

CFA Model Fit Statistics Grades 4-6

Model	LL	Parameters	AIC	BIC	ΔAIC	ΔBIC
1(G2)	-18,699.92	39	37,477.85	37,649.97	_	_
2(G2)	-18,708.49	40	37,496.97	37,673.51	19.12	23.54
3a(G2)	-18,633.92	42	37,351.84	37,537.21	-145.13	-136.3
3b(G2)	-18,694.64	43	37,475.18	37,620.36	123.34	83.15

Metric and Scalar Invariance: MLMG-CFA Model Results

This section reports the results of five competing MLMG-CFA models used to test metric and scalar invariance at multiple levels. The results are displayed in Table 16.

Model 5aH0 (see Figure 2) was the original two-level model 5a with three factors at the between and within level and no estimate for the dispersion parameter. The dispersion parameter was not estimated in MLMG-CFA making this baseline more comparable to other models used in invariance testing. Model H1 maintained equivalence in all parameters of interest but included estimates for a correlated between-level variable that represented rater effect for each group. These between-level effects, indicated in the Mplus input as E1 (rater effect on Group 1) and E2 (rater effect on Group 2) demonstrated that there were only slight differences in the rater effect and those differences were not significant (p=.235).

Models 9aH2 through 9aH5 were done without taking into account variables E1 and E2 due to insufficient evidence that the rater effect differed from one group to the next. Model 9aH2, the configural model, was the least restrictive model of the models used in determining measurement invariance. Factor loadings were freely estimated at both the within and between levels. Intercepts were freely estimated at the between level. Model fit improved from Model 9aH2 to Model 9aH3 (ΔBIC = -60.08). This supports the assumption of factorial invariance (both

weak and strong) at the within level. Model 9aH4 represents the metric (weak factorial) model at both the within and the between level. Factor loadings in this model were equated at both levels. Model fit improved from Model 9aH3 to Model 9aH4 (ΔBIC = -36.99), indicative of weak factorial invariance at the between level in addition to the weak and strong factorial invariance at the within level indicated by the previous comparison. Model 9aH5 was the scalar (strong factorial) model. In this model, factor loadings were equated at the within and between levels. Intercepts at the between level were also equated. Model fit improved from Model 9aH4 to Model 9aH5 (ΔBIC = -100.70). Model 9aH5 was the best fitting model, indicative of scalar (strong factorial) invariance in addition to metric (weak factorial) invariance at both between and within levels.

Table 16

MLMG-CFA Model Fit Statistics

Model	LL	Parameters	AIC	BIC	ΔAIC	ΔBIC
9a H0	-41196.8	45	82,578.63	82,810.06	_	_
9a H1	-39861.6	52	79,827.28	80,094.71	-2,751.35	-2,715.35
9a H2	-39077.8	92	78,399.59	78,812.59	-1,427.69	-1,282.12
9a H3	-39093.2	78	78,342.37	78,743.51	-57.22	-69.08
9a H4	-39096.1	69	78,336.24	78,706.52	-6.13	-36.99
9a H5	-39092.2	59	78,302.40	78,605.82	-33.84	-100.70

Summary

This chapter reported the results of EFA, CFA, MCFA, and MLMG-CFA models used to respond to the three questions that this study addressed. Model fit as indicated by BIC indicated that a three-factor model at both the between and within level was the best fitting model. These results also provided evidence of both weak and strong factorial invariance at both the within and the between levels.

CHAPTER 5

Discussion

Within the framework of classroom quality observation instruments, this study demonstrated and presented procedures for estimating model structure in the presence of categorical and count data, specifically data with Poisson and negative binomial distributions. This study cross-validated a three-factor model structure for 13 behavioral indicators and estimating rater-level effects for the overall and the group levels. In addition, this study tested factorial invariance using MLMG-CFA. This study presented a series of progressively constrained models in which the final model, the strong factorial model, best fit the data, indicating strong factorial invariance.

Factor Structure of the JPAS

This study tested the single-factor, two-factor, and three-factor models using Complex EFA. Change in AIC and BIC indicated that the three-factor model fit best. According to this model, classroom teaching quality as measured by the JPAS can be divided into three distinct factors. This three-factor model was cross-validated by results of CFA and MCFA performed on a separate data set. Results also indicated that the same three-factor model existed at both the within and the between level.

Multiple obstacles unique to count and highly dispersed count data added complexity to modeling the data that does not normally exist when conducting EFA. Count data cannot currently be modeled in a multilevel EFA, making it necessary instead to use a Complex EFA model instead to account for between-level variance. Even within the Complex EFA framework, the dispersion parameter associated with the negative binomial distribution data cannot be currently modeled. Models that include count data do not produce a correlation matrix, rendering

traditional methods of testing EFA models against one another impossible. In the absence of eigenvalues, scree plots, and absolute and relative fit indices, model fit was determined instead by -2LL, AIC, and BIC fit indices alongside factor loading estimates. The many biases that may have been introduced through procedural limitations when modeling count data using complex EFA necessitated a series of CFA models to determine if results could be cross-validated. Results were cross-validated and the three-factor model was established at both the within and the between levels. This should not be an indication that other models are not viable, and further attempts to cross-validate results are encouraged.

Rater-Level Variance

Rater-level variance is known to be a significant source of variance in classroom teaching observation ratings. The rater level variance of this study was much higher than would be expected with an overall rater variance estimated at .677. This is significant because it is indicative that teacher ratings were more dependent on the rater who was observing them than they were on the quality of the classroom instruction that was being observed. This is especially unexpected because the belief not only in the Jordan School District, but in other districts that use the JPAS has been that because count data was being used, resulting ratings would be far more objective than ratings that used a categorical scale. One of the problems, however, is that principals are looking for many behaviors at once, and principals likely vary in their ability to do this. Additionally, principals may differ in what they perceive to be the behavior they are tallying. Regardless of the cause, there was more variance occurring between principals who were observing classroom teaching than there was between teachers who were being observed.

Rater-level variance did not differ significantly from lower to upper grades ($p = .235$). This indicates that the problem of rater-level variance is similar regardless of whether the principal is observing lower grades in the elementary or upper grades in the elementary.

Invariance Across Groups

As far as can be found in the literature, this is the first time MLMG-CFA was conducted using classroom teaching observational data, and the first time that it was conducted using count data. In a field where multiple variables such as student demographics, teacher demographics, grade level, content, and teacher experience level can all have a significant impact on a teacher's classroom instructional quality rating, studies such as this one that test within-level invariance as it occurs in a multilevel structure will help to move the field forward into the development and testing of instruments that are content and grade-level specific. In pulling away from assumptions that guided theories in previous decades, both researchers and practitioners will be able to make better-informed decisions, and researchers will be able to draw less-biased conclusions regarding the relationships between classroom instructional quality and other aspects of teacher evaluation that might be gleaned through stakeholder observations or more sophisticated future measures of student performance.

This is not an instance of one tool being better than the other. Most tools currently in use to measure teachers – be it their content knowledge, the instructional quality they provide, their relationships with students and parents, or their influence on student achievement, are imperfect and prone to bias. Rather than expect that one measure should take precedence over the other. All measures should be refined and reevaluated on a regular basis so that resulting ratings are informative to professional development and the improvement of the profession as a whole.

Limitations

The limitations of this study can be divided into several categories. Computational limitations, instrument design limitations, and scope of study limitations.

Computational Limitations

When conducting EFA and a MLMG-CFA with count data having a negative binomial distribution, a dispersion parameter cannot be estimated as it can when using a CFA or MCFA model. As a result, some of the estimates will be biased as the models do not represent the data well. As noted in comparisons of CFA and MCFA with and without the dispersion parameters, the difference in fit is large: Models that estimate a dispersion parameter fit better than models that do not.

Other computational limitations include the inability to obtain indicator-level ICC in the Mplus program. Muthén and Muthén have explained as recently as 2008, accurate ICCs are not possible because variance/residual variance for count variables cannot be defined. Calculations were done by hand using formulas outlined from research in the field of biology (Nakagawa et al., 2017); however, the accuracy of results are still in question.

Another limitation has to do with the complexity of models in the presence of count data. The more complex the model becomes, the more difficult it is for the average personal computer to manage the processing required when points of integration exceed memory capacity. For this reason, integration had to be limited to a maximum of 10,000 using the input INTEGRATION=MONTECARLO (10,000). The more complex, the higher the chance that with this limitation in place, the program will not reach convergence. In instances where it does, there may still be issues of singularity that may render estimates less accurate. This is probably the

most prohibitive computational aspect of negative binomial count data modeled in multilevel multigroup analysis.

Instrument Design Limitations

A variety of instrument design limitations that are due, in part, to the age of the instrument, prevented every behavioral indicator from inclusion in the model. These design flaws included multiple checklist indicators, indicators that would not be expected in every lesson, repeat indicators, and indicators that could give an advantage to one subject area over another. The inclusion of checklist indicators is inherent to many of the instruments designed in the 1990s and is problematic to the study because they produce data with very little variance. Getting rid of them makes the model easier to estimate, but it does fundamentally change the overall instrument. Making the choice to eliminate such indicators is one that needed to be done with great care and consideration not just for the psychometric impact, but for the impact on the range of behaviors by which a latent trait is estimated. Because indicators were removed, the results are not truly those for the whole instrument, but those behavioral indicators that provided sufficient variance to be included in the analysis.

Scope of Study Limitations

The scope of this study was limited only to teachers of students in Grades 1-6 at the Jordan School District. Conclusions drawn would not be applicable for preschool teachers, kindergarten teachers, and secondary-level teachers. This was an intentional limitation due in part to differences in the structure of secondary vs. elementary education, and in part differences in the way in which classroom teaching is structured for pre-school and kindergarten. In elementary, grade groups are nested within raters. In secondary, grade groups are not nested within raters, but occur at the between level with middle school grade groups having common

raters and high school grade groups having common raters. This would be a different type of analysis and is one that will have to be explored in a separate study. In addition, this study did not take into account the subject that was being taught at the time of observation nor the time of day in which it was taught. These were worthwhile avenues of exploration; however, model complexity at that stage would be prohibitive and likely lead to convergence problems.

Recommendations to the Jordan School District

Recommendations to the Jordan School District regarding modification to the observation portion of the JPAS fall under two categories: indicator-level adjustments and future development. Keeping in mind that the psychometric properties of ratings that result from the use of any observation system require qualitative exploration, recommendations were given with the caveat that conclusions drawn be tempered by an in-depth qualitative examination of district goals and objectives for classroom instructional quality in addition to a thorough examination of the extant literature on classroom instructional frameworks. This in-depth qualitative examination should include multiple stakeholders who represent the different perspectives of those who are impacted by JPAS results. This includes students, parents, teachers, school-level administrators, and representatives from various district-level departments who utilize ratings in order to inform the professional development that they provide throughout the district.

Indicator-Level Recommendations

The three indicator-level recommendations to the Jordan School District included: indicators to omit from the instrument altogether; indicators to review to determine whether they could realistically be expected in any given thirty-minute observation; indicators to review to expand rating options; and domains that might benefit from an increase in sample observed behaviors.

Indicators to Omit. The following actions are recommended to the Jordan School district relating to indicators that should be omitted. First, the Jordan School District should remove any indicators where 95% of teachers received a perfect rating. These indicators take up principal time and take attention away from other indicators that are more valuable in discerning between higher and lower quality of instruction. Second, the new instrument should eliminate behavioral indicators that are easy to measure but have little relevance to student learning. They should instead include behavioral indicators that better represent the qualities of teaching that have the most impact on student learning.

Indicators to Review. The following actions are recommended to the Jordan School District relating to indicators that should be reviewed. First, the Jordan School District should review indicators that are not expected in any given 30-minute lesson. They should also review indicators that have a high variance and are therefore subject to interpretation by principals as to what the behavior should look like. Indicators that are not expected in every 30-minute lesson should be removed. Indicators that have varied interpretations by principals should either be clarified with further explanation and examples, renamed so as to disconnect the associations caused by the phrasing of the indicators or rubric, or considered for removal if the interpretation remains so broad as to render the indicator more subject to the rater than it is teaching quality.

Indicators to Expand. The following actions are recommended to the Jordan School District regarding category response expansion. The Jordan School District should expand yes and no dichotomous response categories to include four categorical rater responses: (a) not effective, (b) minimally effective, (c) effective, and (d) highly effective. This response structure not only matches what is done in the overall evaluation system, but it is one that principals and teachers are now familiar enough with to use without difficulty. The reason for this

recommendation is to increase the capacity of each categorical indicator to differentiate between varying levels of quality within each indicator. This modification would increase granularity of the data and add more information to the appraisal system as a whole.

Recommendation for Future Development

The following actions are recommended to the Jordan School District for future development of the classroom teaching observation instrument used in the JPAS. The Jordan School District should form a new committee to explore the most recent literature on classroom instructional quality and compile a new literature review that takes into account developments that have occurred since the last thorough review was conducted. It is suggested that members of this committee conduct interviews of teachers, administrators, district personnel, and other key stakeholders in order to obtain input on strengths and weaknesses of the current system so that portions that have high face validity and strong theoretical underpinnings may be maintained. In addition, it is suggested that experts in the field of classroom instructional quality both within and outside of the district be consulted as was done during the initial development of the JPAS. The information gleaned from a thorough exploration of theory combined with information from this and future studies of the psychometric properties of the JPAS in conjunction with a clear vision of district instructional goals will not only increase the validity of using ratings to make determinations as to what aspects of classroom instructional quality are stronger or weaker in each classroom and school, but also improve and expand the use of JPAS ratings as a means to inform professional development and improve classroom instructional quality across the district.

Several avenues of study have potential to add to what has been explored in this study. First, a study of the factor structure of the JPAS at the secondary level should be explored. In addition, at the secondary level, the degree to which content taught impacts the factor structure

of JPAS data need to be explored in order to determine whether separate content-specific evaluations should be developed. A study employing Indicator Response Theory (IRT) would be beneficial in further exploring measurement invariance through an analysis of Differential Indicator Function (DIF). It would also be beneficial to study other variables that might have an impact on the factor structure of the JPAS at the within level such as student gender, race, and socioeconomic status.

One assumption that needs to be tested with regards to count data is the assumption that more is better. Currently, cut scores are established for some count indicators with zero established as the lowest rating with ratings that increase the more tallies a teacher accrues during the observation time. Theoretically, this is an inaccurate way to create cut scores for count indicators. For example, asking 60 factual questions during a 30-minute time frame could easily allow this strategy, which requires less critical thinking on the part of the student, to take up the majority of the instruction time. In this regard, a study that establishes thresholds for indicators would be beneficial. Such a study would establish poor, good, better, and best count tallies. While on some indicators, those counts might peaks at a mid-point and then tapers off in both directions away from the ideal, on other indicators the ratings might follow the standard increased rating with increased count

Finally, an exploration to determine if the current equations established for estimating indicator-level ICC are sufficient or should be improved would be beneficial as the estimates from this study appeared to be very high and may be inaccurate.

Recommendations to Software Developers

It would be beneficial, in the future, to have the ability to estimate a dispersion parameter when conducting MLMG-CFA. This is currently one of the greatest shortcomings of this study.

Having to return to a model that is already significantly biased due to estimating for a Poisson distribution when strong empirical evidence for a negative binomial distribution exists creates a situation where invariance testing may be flawed at both the within and between levels. Adding this feature to Mplus would greatly enhance the ability of researchers to carefully address the unique qualities of highly dispersed data while testing for group invariance at the within level in multilevel SEMs.

Conclusions

This study sought to determine what factor structure best represented the underlying relationship between the JPAS behavioral indicators of classroom instructional quality in Grades 1-6. Secondarily, this study also sought to estimate the impact that the variables of rater and grade level had on those relationships. The answers to these questions do not represent a singular directive for school district personnel, but they do serve to add valuable information to the decision making process in addition to introducing new questions that should be explored prior to revising or replacing the current instrument. Most important, results from the ICC give school district personnel strong empirical evidence for making changes in the instrument in order to reduce the impact of the rater on teacher ratings.

The results of this study also add to the body of research on classroom observation instruments by introducing rigorous methods of establishing the factor structure of classroom observation instruments and testing for invariance across groups when observation data is nested within raters or hierarchical systems such as schools or districts. Specifically, this study introduced methodologies that will allow researchers to create models that examine variables of interest such as grade level or demographic variables that exist at the within level. Additionally, this study introduces rigorous methods of addressing CFA, MCFA, and MLMG-CFA when data

is discrete and does not follow a normal Gaussian dispersion, as is the case for Poisson and negative binomial distributions. While this is one study of a locally developed classroom observation instrument used within a very specific demographic context, the vision of such a study is that it will further an already progressing trajectory towards the use of robust methodologies as a companion to qualitative studies in classroom instructional quality and open up avenues of exploration that lead to increasingly valid and reliable ratings from measures of classroom instructional quality.

REFERENCES

American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*(8), 448-452. https://doi.org/10.3102/0013189X15618385

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Asparouhov, T., & Muthén, B. (2012, November 17). *Multiple group multilevel analysis*. Mplus web notes: No. 16. http://www.statmodel.com/examples/webnotes/webnotes16.pdf

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Press.

Beem, A. L., & Brugman, D. (1985). An exploration of the structure of classroom behavior during values development lessons. *Studies in Educational Evaluation, 11*(3), 339-357. https://doi.org/10.1016/0191-491X(85)90017-3

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2-3), 62-87. https://doi.org/10.1080/10627197.2012.715014

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 8*(4), 42-51. https://doi.org/10.1111/j.1745-3992.2009.00161.x

Briggs, B. A., & Dickerscheid, J. D. (1985). Personality characteristics and preschool teaching behaviors in student teachers. *Educational and Psychological Research, 5*(1), 55–63.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Capie, W., Ellett, C., & Johnson, C. (1980, March). *Relating pupil achievement gains to ratings of secondary student teacher performance* [Paper presentation]. Annual conference of the Eastern Educational Research Association, Norfolk, VA, United States.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311–337. https://doi.org/10.1177/0013164414539163

Charalambous, Y., Kyriakides, E, Leonidas, K., & Tsangaridou, N. (2019). Are teachers consistently effective across subject matters? Revisiting the issue of differential teacher effectiveness. *School Effectiveness and School Improvement, 20*(4), 353-379. https://doi.org/10.1080/09243453.2019.1618877

Chauvin, S. W., Loup, K. S., & Ellett, C. D. (1991, April 4-6). *Development and validation of a comprehensive assessment system for teaching and learning* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Chicago, IL, United States.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom teaching observations. *Educational Researcher*, *45*(6), 378-387. https://doi.org/10.3102/0013189X16659442

Cooley, W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis, 2*(1), 7-25. https://doi.org/10.3102%2F01623737002001007

Crocker, R. K., & Brooker, G. M. (1986). Classroom control and student outcomes in Grades 2 and 5. *American Educational Research Journal, 23*(1), 1–11. https://doi.org/10.3102/00028312023001001

Crawford, A. D., Zucker, T. A., Williams, J. M., Bhavsar, V., & Landry, S. H. (2013). Initial validation of the prekindergarten classroom observation tool and goal setting system for data-based coaching. *School Psychology Quarterly, 28*(4), 277-300. https://doi.org/10.1037/spq0000033

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., Hamre, B., & Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade. *The Elementary School Journal, 112*(1), 16–37. https://doi.org/10.1086/660682

Darling-Hammond, L. (1996). The right to learn and the advancement of teaching: Research, policy, and practice for democratic education. *Educational Researcher, 25*(6), 5-17. https://doi.org/10.3102/0013189X025006005

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Association for Supervision and Curriculum Development.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Association for Supervision and Curriculum Development.

Dorety, K. M., & Jacobs, S. (2015). *State of the states 2015: Evaluating teaching, leading and learning*. National Council on Teacher Quality. https://www.nctq.org/publications/State-of-the-States-2015:-Evaluating-Teaching,-Leading-and-Learning

Downer, J. T., López, M. L., Grimm, K. J., Hamagami, A., Pianta, R. C., & Howes, C. (2012). Observations of teacher-child interactions in classrooms serving Latinos and dual language learners: Applicability of the Classroom Assessment Scoring System in diverse settings. *Early Childhood Research Quarterly*, *27*(1), 21-32. https://doi.org/10.1016/j.ecresq.2011.07.005

Doyle, W. (1986). Classroom organization and management. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 392-432). Macmillan.

Dunteman, G. H. (1989). *Quantitative applications in the social sciences: Principal components analysis*. SAGE Publications. https://doi: 10.4135/9781412985475

Ellett, C., Loup, K., & Chauvin, S. (1991). Development, validity and reliability of a new generation of assessments of effective teaching and learning: Future directions for the study of learning environments. *The Journal of Classroom Interaction, 26*(2), 25-39.

Enders, C. K., (2010). *Applied missing data analysis*. Guilford Press. https://doi.org/10.1111/j.1467-842X.2012.00656.x

Every Student Succeeds Act of 2015. Pub. L. No.114-195 § 114 Stat. 1177 (2015). Retrieved from https://www.congress.gov/bill/114th-congress/senate-bill/1177/text.

Evertson, C. M., Anderson, C., Anderson, L. M., & Brophy, J. (1980). Relationships between classroom behaviors and student outcomes in junior high mathematics and English classes. *American Educational Research Journal, 17*(1), 43-60. https://doi.org/ 10.2307/1162507

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), 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford Press.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology, 39*(2), 291-314. https://doi.org/10.1111/j.1744-6570.1986.tb00583.x

Gagné, P., & Hancock G. R., (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*(1), 65–83. https://doi.org/10.1207/s15327906mbr4101_5

Gill, B., Shoji, M., Coen, T., & Place, K. (2016, November). *The content, predictive power, and potential bias in five widely used teacher observation instruments*. Institution of Education Sciences. https://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=4474

Gitomer, D., Bell, C., Qi, Y., Croft, A., Leusner, D. M., McCaffrey, D., & Pianta, R. (2014). Improving observational score quality: Challenges in observer thinking. In K. Kerr, R. Pianta, & T. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 50–97). Jossey-Bass. https://doi.org/10.1002/9781119210856.ch3

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* National Comprehensive Center for Teacher Quality. https://doi.org/10.3102/0013189X16659442

Gorsuch R. L. (1990). Common factor analysis versus component analysis: Some well and little known facts. *Multivariate Behavioral Research, 25*(1), 33-39. https://doi.org/10.1207/s15327906mbr2501_3

Grossman, P., Cohen, J., & Brown, L. (2014). Understanding instructional quality in English Language Arts: Variations in the relationship between PLATO and value-added by

content and context. In Kerr, K., Pianta, R., & Kane, T. (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 303-331). Jossey-Bass. https://doi.org/10.1002/9781119210856.ch10

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445–470.

Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal*, *45*(1), 184–205. https://doi.org/10.3386/w16015

Hafen, C., Ruzek, E., Gregory, A., Allen, J., & Mikami, A. (2015). Focusing on teacher-student interactions eliminates the negative impact of students' disruptive behavior on teacher perceptions. *International Journal of Behavioral Development, 39*(5), 426-431. https://doi.org/10.1177/0165025415579455.

Havik, T., & Westergård, E. (2020). Do teachers matter? Students' perceptions of classroom interactions and student engagement. *Scandinavian Journal of Educational Research. 64*(4), 488-507. https://doi.org/10.1080/00313831.2019.1577754.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the Mathematical Quality of Instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430-511.

Holmes, F., & Bolin, J. (2017). *Multilevel modeling using Mplus*. CRC Press.

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structure Equation Modeling*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling, 21*(1), 31-39. https://doi.org/10.1080/10705511.2014.856694

Jensen, B., Wallace, T. L., Steinberg, M. P., Gabriel, R. E., Dietiker, L., Davis, D. S., Kelcey, B., Minor, E. C., Halpin, P., & Rui, N. (2019). Complexity and scale in teaching effectiveness research: Reflections from the MET Study. *Education Policy Analysis Archives, 27*(7). https://doi.org/10.14507/epaa.27.3923

Jordan School District. (1993). *Jordan Performance Appraisal System domains document.*

Jordan School District. (1996). *Jordan Performance Appraisal System development manual*.

Kallison, J. M., Jr. (1986). Effect of lesson organization on achievement. *American Educational Research Journal*, *23*(2), 337-347. https://doi.org/10.2307/1162963

Kim, E., Kwok, O., & Yoon, M. (2012). Testing factorial invariance in multilevel data: A Monte Carlo study. *Structural Equation Modeling*, *19*(2), 250- 267. https://doi.org/10.1080/10705511.2012.659623

Kim, E., Yoon, M., Wen, Y., Luo, W., & Kwok, O. (2015). Within-level group factorial invariance with multilevel data: Multilevel factor mixture and multilevel MIMIC models. *Structural Equation Modeling*, 22, 603-616. doi.org/10.1080/10705511.2014.938217

Koch, G. G. (2006). Intraclass correlation coefficient. *Encyclopedia of statistical sciences.* American Cancer Society. https://doi.org/10.1002/0471667196.ess1275.pub2

Lockwood, J. R., & McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy*, *4*(4), 439-467. https://doi.org/10.1162/edfp.2009.4.4.439

Malmberg, L., Hagger, H., Burn, K. Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology, 102*(4), 916-932. https//doi.org/ 10.1037/a0020920

Manaf, Z. A. (1995). Model testing through confirmatory factor analysis using LISREL: An explication of the methodology using secondary data. *Research in Education*, *53*(1), 11–23. https://doi.org/10.1177/003452379505300102

Mashburn, A. J., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2006). Teacher and classroom characteristics associated with teachers' ratings of prekindergartners' relationships and behaviors. *Journal of Psychoeducational Assessment, 24*(4), 367-380. https://doi.org/10.1177/0734282906290594

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015) Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement Issues and Practice, 34*(2), 34-46.

McNemar, Q. (1951). The factors in factoring behavior. *Psychometrika, 16,* 353–359. https://doi.org/10.1007/BF02288800

Mikeska, J. N., Holtzman, S., McCaffrey, D. F., Liu, S., & Shattuck, T. (2019). Using classroom observations to evaluate science teaching: Implications of lesson sampling for measuring science teaching effectiveness across lesson types. *Science Education*, *103*(1), 123-144. https://doi.org/10.1002/sce.21482

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* Routledge.

Muthén, B., & Muthén L. K. (2008, July 16). *Intraclass correlation.* Mplus discussion. http://www.statmodel.com/discussion/messages/12/18.html

Muthén, B., & Muthén L. K. (2010, February 27). *Latent variable mixture modeling.* Mplus discussion. http://www.statmodel.com/discussion/messages/13/5162.html?1582075371

Muthén, B., & Muthén L. K. (2017). *Mplus user's guide* (8th Ed.). Muthén & Muthén. http://www.statmodel.com/ugexcerpts.shtml

Muthén, B., & Muthén L. K. (2020). *Confirmatory Factor Analysis.* Mplus discussion. http://www.statmodel.com/discussion/messages/9/9.html

Nakagawa, S. Johnson, P. C. D., & Schielzeth, H. (2017). Coefficient of determination R^2 and intra-class correlation coefficient ICC from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*(134), 1-11. https://doi.org/10.1098/rsif.2017.0213

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* https://www2.ed.gov/pubs/NatAtRisk/risk.html

No Child Left Behind Act of 2001: Qualifications for teachers and professionals, 20 U.S.C. § 6319 (2008).

Pakarinen, E., Kiuru, N., Lerkkanen, M.-K., Poikkeus, A.-M., Siekkinen, M., & Nurmi, J. E. (2010). Classroom organization and teacher stress predict learning motivation in kindergarten children. *European Journal of Psychology of Education, 25*(3), 281–300. https://doi.org/10.1007/s10212-010-0025-6

Pianta, R. C., Hamre, B. K., Haynes, N. J, Mintz, S., & La Paro, K. M. (2006). *CLASS classroom assessment scoring system manual: Middle secondary version pilot*. Teachstone.

Pianta, R. C. La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System. manual Pre-K.* Brookes.

Pilburn, M., & Sawada, D. (2000). *Reformed Teaching Observation Protocol (RTOP) reference manual. NSF technical report*. Arizona State University.

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, *121*(2), 183-212. https://doi.org/10.1086/679390

Qi, Y., Bell, C., & Gitomer, D. (2014, April 3-7). *The role of topic and activity structure in teacher observation scores* [Paper presentatioon]. Annual meeting of the American Educational Research Association, Philadelphia, PA, United States.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge. https://doi.org/10.1111/j.1751-5823.2011.00149_24.x

Rosenshine, B. (1983). Teaching functions in instructional programs. *The Elementary School Journal, 83*(4), 335-351. https://doi.org/10.1086/461321

Ryu, E. (2015). Multiple group analysis in multilevel structural equation model across level 1 groups. *Multivariate Behavioral Research*, *50*(3), 300-315. https://doi.org/10.1080/00273171.2014.1003769

Ryu, E., & Mehta, P. (2017). Multilevel factorial invariance in n-level structural equation modeling (nSEM). *Structural Equation Modeling, 24*(6), 936-959. https://doi.org/10.1080/10705511.2017.1324311

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299-311. https://doi.org/10.1007/BF00973726

Sandilos, L. E., & DiPerna, J. C. (2014). Measuring quality in kindergarten classrooms: Structural analysis of the Classroom Assessment Scoring System (CLASS K-3). *Early Education and Development, 25*(6), 894-914. https://doi.org/10.1080/10409289.2014.883588

Sandilos, L. E., Wollersheim S., DiPerna, J. Lei, P. W., & Cheng, W. (2016). Structural validity of CLASS K-3 in primary grades: Testing alternative models. *School Psychology Quarterly. 32*(2), 226-239. https://doi.org/10.1037/spq0000155

Sirotnik, K. A. (1980). Psychometric implications of the unit-of-analysis problem (with examples from the measurement of organizational climate). *Journal of Educational Measurement, 17*(4), 245-282. https://doi.org/10.1111/j.1745-3984.1980.tb00831.x

Sloat, E., Amrein-Beardsley, A., & Sabo, K. E. (2017). Examining the factor structure underlying the TAP System for Teacher and Student Advancement. *AERA Open, 3*(4), 1-18. https://doi.org/10.1177/2332858417735526

Spearman, C. (1904). 'General intelligence,' objectively determined and measured. *The American Journal of Psychology, 15*(2), 201–293. https://doi.org/10.2307/1412107

Spearman, C. (1927). *The abilities of man.* Macmillan.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1-10.

Virtanen, T., Pakarinen, E., Lerkkanen, M. K., Poikkeus, A.M., Siekkinen, M., & Nurmi, J. E. (2018). A validation study of Classroom Assessment Scoring System-Secondary in the Finnish School Context. *Journal of Early Adolescence, 38*(6), 849-880. doi:10.1177/0272431617699944

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness* (2nd Ed.). The New Teacher Project.

Weinstein, C. S. (1979). The physical environment of the school: A review of the research. *Review of Educational Research, 49*(4), 577-610. https://doi.org/10.2307/1169986

APPENDIX A

Domain I

Table A1

Indicators from Domain I: Classroom management

Indicator	Label	Description	Data Type
1	Students off-task	A tally of the number of students off task during observation.	Count
2	Interrupts or Obscures Instruction	A tally mark is recorded for each time a teacher interrupts or obscures instruction.	Count
3	Fails to address misunderstandings	A tally mark is recorded for each time the teacher does not use an opportunity to address a student concern or misunderstanding.	Count
4	Fails to respond immediately to disruptive behavior	A tally mark is recorded each time a teacher does not recognize and remedy disruptive behavior.	Count
5	Adjusts instruction	Determines whether or not the teacher adjusts instruction to meet the needs of diverse learners.	Categorical
6	Smooth transitions	Determines whether or not there are disruptions during transitions. Three responses are available: yes, no, no transitions.	Categorical
7	Positive learning climate	The teacher engages students in a positive and inclusive manner or not.	Categorical
8	Responds consistently to behaviors	The teacher is consistent in their response to student behaviors.	Categorical
9	Applies low-key tactics for misbehavior	Low-key tactics for misbehavior are used effectively.	Categorical
10	Identifies initiators of disruptive behavior	The teacher identifies those who are initiating the disruptions in order to stop them quickly.	Categorical

Table continues on next page

Table A1 (Continued)

Indicator	Label	Description	Data Type
11	Use of management routines	Determines the degree to which classroom routines are outlined and followed.	Categorical
12	Classroom management	Determines the degree to which the teacher uses differentiated and effective classroom management techniques.	Categorical
13	Minutes of nonacademic time	A tally mark is recorded for each minute lost to nonacademic activities.	Count

APPENDIX B

Domain II

Table B1

Indicators from domain II - Delivering instruction

Indicator	Label	Description	Data Type
14	Factual questions	Tally each time the teacher asks factual questions to assess learning.	Count
15	Explains academic concepts	Tally each time a teacher explains an academic concept.	Count
16	Demonstrates skills/procedures	Tally each time the teacher models a skill or procedure or uses manipulatives, visual representations, or hands-on material to demonstrate a skill or procedure that students are expected to perform.	Count
17	Illustrates relationships	Tally recorded each time the teacher illustrates a relationship by tying new information to concepts students understand.	Count
18	Emphasizes important points	Tally each time the teacher emphasizes an important point in the lesson.	Count
19	Reviews	Tally recorded each time the teacher reviews or summarizes a concept or skill from a previous or current lesson.	Count
20	Pre-assessment	Determines whether or not a teacher has taken the time to determine if the students are prepared with proper skills and/or knowledge for understanding new concepts, materials, or tasks before intruding them.	Categorical
21	Advance Organizer	Determines whether or not a teacher provided an overview of the lesson that helps students prepare for what they'll be learning.	Categorical
22	Teaching/learning strategies	Determines whether or not the teacher used learning strategies such as questioning, study guides, graphic organizers, etc. to help students gain and process new information.	Categorical

Table continues on next page

Table B1 (Continued)

Indicator	Label	Description	Data Type
24	Energy and enthusiasm	Degree to which a teacher displays clearly discernable interest in the subject matter through speech and body language.	Categorical
25	Goals, objectives, and expectations	Determines the degree to which the teacher states goals, objectives, and expectations and relates them to the learning activity.	Categorical
26	Instructional delivery	Determines the degree to which a teacher helps to deepen student understanding by helping them evaluate, create, and think critically about content.	Categorical
27	High-order questions	Tally for each time the teacher incorporates higher-level-thinking questions.	Count
28	Wait time	A tally is recorded each time the teacher asks a question and pauses for at least three seconds before calling on a student.	Count
29	Sustains interactions	Tally for each time the teacher sustains dialogue with a student by asking follow-up questions.	Count
30	Task-oriented peer interaction	Determines if a teacher initiates whole class learning tasks or provides time for academic interaction between students.	Categorical
31	Problem solving	Teacher uses instructional strategies requiring higher-order thinking skills.	Categorical
32	Cause-effect analysis	Determines whether or not a teacher helps students critically think about the subject they are learning using cause-effect analysis.	Categorical
33	Authentic learning experience	Determines if a teacher helps students practically apply what they have learned.	Categorical
34	Brainstorming and use of ideas	Determines whether or not a teacher helps students to brainstorm and develop multiple ideas regarding the content being learned.	Categorical
35	Prepares students for activities	Degree to which a teacher prepares students for activities using directions and ensuring students understand them.	Categorical

Table continues on next page

Table B1 (Continued)

Indicator	Label	Description	Data Type
36	Supervises independent practice	Determines the degree to which a teacher walks around the room helping students as needed with independent practice.	Categorical
37	Correctives	Determines the degree to which a teacher responds to incorrect responses by rephrasing questions, providing prompts or briefly re-teaching material.	Categorical
38	Monitors student performance	Determines the degree to which a teacher monitors and guides all students in their learning in order to help them increase their level of performance and understanding.	Categorical

APPENDIX C

Domain III

Table C1

Indicators From Domain III - Interacting With Students

Indicator	Label	Description	Data Type
39	Student participation	Tally is recorded each time the teacher initiates an interaction with a different student about the content.	Count
40	Academic feedback	Tally is recorded each time the teacher provides academic feedback.	Count
41	Gets student attention	Tally for each time the teacher uses a procedure to get student attention before moving forward in the lesson.	Count
42	Encourages reluctant students	Tally is recorded each time the teacher recognizes a student who is not participating and involves them.	Count
43	Reinforces desired behavior	Tally is recorded if the teacher offers specific praise to students.	Count
44	Acknowledges learning efforts	Tally is recorded for each statement or nonverbal gesture a teacher makes to acknowledge the effort a student has made learning new material.	Count
45	Student demonstrations of knowledge or skills	Students are given time to show their knowledge or skills with others.	Categorical
46	Practices communication skills	Teacher teaches reading, writing, listening, and speaking skills for effective communication.	Categorical
47	Guided practice	Teacher provides guided practice for new concepts, tasks, or procedures.	Categorical
48	Checks for understanding	Teacher checks periodically for understanding of information being presented.	Categorical
49	Learning environment	Degree to which a teacher is engaged with/engages each of their students in academic learning.	Categorical

APPENDIX D

Reasons Indicators Were Not Retained

Table D1

JPAS Indicators: Reasons Indicators Were Not Retained

Indicator	Description	Reason for Removal
1	Number of students off task.	Time dependent
2	Teacher interrupts of obscures instruction.	Insufficient variance
3	Teacher fails to address misunderstandings.	Insufficient variance
4	Teacher does not remedy misbehavior.	Insufficient variance
5	Teacher adjusts instruction.	Insufficient variance
6	Disruptions during transitions.	Insufficient variance
7	Teacher engages students in inclusive manner.	Insufficient variance
8	Teacher is consistent in response to behaviors.	Insufficient variance
12	Differentiated and effective classroom management.	Multiple behaviors
13	Number of minutes lost to academic time.	Insufficient variance
16	Uses manipulatives, visual representations, or hands-on material to demonstrate a skill.	Multiple behaviors
19	Reviews or summarizes a concept or skill from a previous or current lesson.	Multiple behaviors
20	Prepares for new skills and/or knowledge for understanding new concepts, materials, or tasks.	Multiple behaviors
21	Provides a brief overview of the lesson.	Lesson dependent
22	Uses learning strategies such as questioning, study guides, graphic organizers, etc. in order to help students gain and process new information.	Insufficient variance
23	Structure & sequence of lessons helps students master skills and understanding.	Insufficient variance

Table continues on next page

Table D1 (Continued)

Indicator	Description	Reason for Removal
30	Initiates whole class learning tasks or provides time for academic interaction between students.	Multiple behaviors
32	Determines whether or not a teacher helps students critically think about the subject they are learning using cause-effect analysis.	Insufficient variance
33	Determines whether or not a teacher helps students practically apply what they have learned.	Lesson dependent
34	Determines whether or not a teacher helps students to brainstorm and develop multiple ideas regarding the content being learned.	Lesson dependent
36	Determines the degree to which a teacher walks around the room helping students as needed with independent practice.	Lesson dependent
37	Determines the degree to which a teacher responds to incorrect responses by rephrasing questions, providing prompts or briefly re-teaching material.	Multiple behaviors
45	Determines whether or not students are given an opportunity to demonstrate their knowledge or skills with others	Lesson dependent
46	Determines whether or not the teacher teaches reading, writing, listening, and speaking skills for effective communication.	Multiple behaviors
47	Determines whether or not a teacher provides guided practice for new concepts, tasks, or procedures.	Lesson dependent
48	Determines whether or not a teacher checks periodically for understanding of information being presented.	Insufficient variance
49	Determines the degree to which a teacher is engaged with and engages each of their students in academic learning. Includes explanation, discussion, review, reading aloud etc.	Multiple behaviors

APPENDIX E

Cluster-Level Mean by Indicator

Table E1

Cluster-Level Mean by Indicator

	Indicators												
Cluster	i14	i15	i17	i18	i27	i28	i29	i39	i40	i41	i42	i43	i44
1	22.38	0.78	2.16	1.94	6.81	22.38	0.78	2.16	1.94	6.81	2.16	1.94	6.81
2	34.88	0.79	8.15	9.94	13.44	34.88	0.79	8.15	9.94	13.44	8.15	9.94	13.44
3	19.44	0.94	1.61	2.39	6.10	19.44	0.94	1.61	2.39	6.10	1.61	2.39	6.10
4	19.72	0.94	1.17	2.42	6.06	19.72	0.94	1.17	2.42	6.06	1.17	2.42	6.06
5	32.31	0.96	1.86	1.08	9.05	32.31	0.96	1.86	1.08	9.05	1.86	1.08	9.05
6	22.84	1.03	1.50	2.59	6.99	22.84	1.03	1.50	2.59	6.99	1.50	2.59	6.99
7	15.52	1.03	1.42	1.52	4.87	15.52	1.03	1.42	1.52	4.87	1.42	1.52	4.87
8	22.91	1.06	3.13	2.31	7.35	22.91	1.06	3.13	2.31	7.35	3.13	2.31	7.35
9	28.92	1.21	3.67	1.63	8.85	28.92	1.21	3.67	1.63	8.85	3.67	1.63	8.85
10	37.75	1.21	0.71	4.42	11.02	37.75	1.21	0.71	4.42	11.02	0.71	4.42	11.02
11	31.67	1.22	0.69	0.28	8.47	31.67	1.22	0.69	0.28	8.47	0.69	0.28	8.47
12	31.91	1.24	2.29	3.97	9.85	31.91	1.24	2.29	3.97	9.85	2.29	3.97	9.85
13	18.00	1.25	0.50	1.44	5.30	18.00	1.25	0.50	1.44	5.30	0.50	1.44	5.30
14	25.13	1.27	1.60	3.77	7.94	25.13	1.27	1.60	3.77	7.94	1.60	3.77	7.94
15	26.75	1.28	1.67	1.08	7.69	26.75	1.28	1.67	1.08	7.69	1.67	1.08	7.69
16	20.20	1.30	0.77	0.97	5.81	20.20	1.30	0.77	0.97	5.81	0.77	0.97	5.81
17	29.62	1.38	2.92	4.38	9.58	29.62	1.38	2.92	4.38	9.58	2.92	4.38	9.58
18	39.14	1.43	1.71	1.21	10.88	39.14	1.43	1.71	1.21	10.88	1.71	1.21	10.88
19	20.13	1.43	1.09	10.13	8.20	20.13	1.43	1.09	10.13	8.20	1.09	10.13	8.20
20	22.17	1.54	1.08	5.79	7.65	22.17	1.54	1.08	5.79	7.65	1.08	5.79	7.65
21	38.71	1.56	2.32	4.29	11.72	38.71	1.56	2.32	4.29	11.72	2.32	4.29	11.72
22	30.33	1.60	0.93	2.03	8.73	30.33	1.60	0.93	2.03	8.73	0.93	2.03	8.73
23	35.21	1.63	1.04	1.88	9.94	35.21	1.63	1.04	1.88	9.94	1.04	1.88	9.94
24	22.64	1.64	1.57	5.79	7.91	22.64	1.64	1.57	5.79	7.91	1.57	5.79	7.91
25	25.19	1.75	1.94	2.69	7.89	25.19	1.75	1.94	2.69	7.89	1.94	2.69	7.89
26	22.66	1.78	1.53	4.00	7.49	22.66	1.78	1.53	4.00	7.49	1.53	4.00	7.49
27	35.95	1.85	1.70	2.10	10.40	35.95	1.85	1.70	2.10	10.40	1.70	2.10	10.40
28	18.38	1.85	0.47	3.06	5.94	18.38	1.85	0.47	3.06	5.94	0.47	3.06	5.94
29	22.93	1.86	4.00	4.71	8.38	22.93	1.86	4.00	4.71	8.38	4.00	4.71	8.38
30	33.07	1.86	3.00	2.86	10.20	33.07	1.86	3.00	2.86	10.20	3.00	2.86	10.20

Table continues on next page

Table E1 (Continued)

	Indicators												
Cluster	i14	i15	i17	i18	i27	i28	i29	i39	i40	i41	i42	i43	i44
31	27.00	2.00	2.00	1.50	8.13	27.00	2.00	2.00	1.50	8.13	2.00	1.50	8.13
32	29.71	2.04	4.36	3.16	9.82	29.71	2.04	4.36	3.16	9.82	4.36	3.16	9.82
33	16.47	2.08	3.74	3.61	6.47	16.47	2.08	3.74	3.61	6.47	3.74	3.61	6.47
34	30.56	2.09	2.09	0.62	8.84	30.56	2.09	2.09	0.62	8.84	2.09	0.62	8.84
35	21.29	2.17	1.13	2.46	6.76	21.29	2.17	1.13	2.46	6.76	1.13	2.46	6.76
36	19.00	2.25	2.33	2.00	6.40	19.00	2.25	2.33	2.00	6.40	2.33	2.00	6.40
37	21.75	2.38	1.56	5.63	7.83	21.75	2.38	1.56	5.63	7.83	1.56	5.63	7.83
38	24.43	2.43	1.17	1.14	7.29	24.43	2.43	1.17	1.14	7.29	1.17	1.14	7.29
39	31.00	2.60	1.60	0.72	8.98	31.00	2.60	1.60	0.72	8.98	1.60	0.72	8.98
40	17.67	2.61	1.61	2.00	5.97	17.67	2.61	1.61	2.00	5.97	1.61	2.00	5.97
41	30.77	2.77	3.00	3.82	10.09	30.77	2.77	3.00	3.82	10.09	3.00	3.82	10.09
42	25.83	2.83	1.33	1.50	7.88	25.83	2.83	1.33	1.50	7.88	1.33	1.50	7.88
43	36.93	3.11	1.68	3.18	11.22	36.93	3.11	1.68	3.18	11.22	1.68	3.18	11.22
44	22.80	3.12	1.60	3.48	7.75	22.80	3.12	1.60	3.48	7.75	1.60	3.48	7.75
45	22.56	3.13	3.38	3.19	8.06	22.56	3.13	3.38	3.19	8.06	3.38	3.19	8.06
46	40.15	3.33	1.00	2.79	11.82	40.15	3.33	1.00	2.79	11.82	1.00	2.79	11.82
47	22.39	3.56	2.22	8.28	9.11	22.39	3.56	2.22	8.28	9.11	2.22	8.28	9.11
48	33.92	3.75	0.67	2.33	10.17	33.92	3.75	0.67	2.33	10.17	0.67	2.33	10.17
49	27.54	4.77	0.96	1.73	8.75	27.54	4.77	0.96	1.73	8.75	0.96	1.73	8.75
50	32.88	5.56	2.40	2.64	10.87	32.88	5.56	2.40	2.64	10.87	2.40	2.64	10.87
51	24.78	6.50	3.97	11.19	11.61	24.78	6.50	3.97	11.19	11.61	3.97	11.19	11.61

APPENDIX F

Mplus Input Model 5a H0

```
INPUT INSTRUCTIONS
TITLE: JPAS ELEMENTARY MLMG CFA
DATA: FILE = C:\JPAS2.txt;
VARIABLE: NAMES =
        obsnum evalid id teach rater schid lev grade studcc prov
        sub classtype ampm obstime classtim
        grtime indtime disr disr2 disr3
        i1-i49;
USEVARIABLES=
        rater
        i17 i18 i41-i44
        i14 i40
        i15 i27-i29 i39;
        COUNT =
        i17 i18 i41-i44(nb)
        i14 i40 (nb)
        i15 i27-i29 i39(nb);
        CLASSES = c(2);
KNOWNCLASS =
        c(lev = 0 1); !grouping is  by grade level grouping 1-3 and 4-6
CLUSTER =
        rater; !clustering is by rater
MISSING =
        ALL (999); ! missing data is defined by the number 999
ANALYSIS:
TYPE = TWOLEVEL MIXTURE;
ESTIMATOR = MLF;
INTEGRATION=MONTECARLO(500);
    mcon = .1
MODEL:
    %WITHIN%
    %OVERALL% !Comparison Group factors equated with between-level factors
    wF1 by i14@1
    i40* (lam40); !Factor 1
    wF2 by i17@1
    i18* (lam18) !Factor 2
    i41-i44* (lam41-lam44);
    wF3 by i15@1
    i27-i29* (lam27-lam29)
    i39* (lam39); !Factor 3
    %C#1% !Group#1 Grades 1-3
```

```
wF1*1
wF2*1
wF3*1;
%C#2% !Group #2 Grades 4-6
wF1*1
wF2*1
wF3*1;

%BETWEEN%
%OVERALL%
bF1 by i14@1
i40* (lam40); !Factor 1
bF2 by i17@1
i18* (lam18) !Factor 2
i41-i44* (lam41-lam44);
bF3 by i15@1
i27-i29* (lam27-lam29)
i39* (lam39); !Factor 3
e1 by bF1@0 bF2@0 bF3@0;
e2 by bF1@0 bF2@0 bF3@0;
bF1@0 bF2@0 bF3@0;
[e1@0 e2@0]; e1*1; e2*1; e1 with e2*0;
%C#1% ! Group #1 Grades 1-3
[bF1@0]
[bF2@0]
[bF3@0]; e1 by bF1@1 bF2@1 bF3@1;
%C#2% ! Group #2 Grades 4-6
[bF1*0]
[bF2*0]
[bF3*0]; e2 by bF1@1 bF2@1 bF3@1;
```