2020

Student Voices in Teacher Evaluation: A Multilevel, Latent Factor Investigation of Teacher Quality

Kathleen J. Hoff
Walden University

Walden University

College of Social and Behavioral Sciences

This is to certify that the doctoral dissertation by

Kathleen J. Hoff

has been found to be complete and satisfactory in all respects,
and that any and all revisions required by
the review committee have been made.

Review Committee
Dr. Steven Kriska, Committee Chairperson, Psychology Faculty
Dr. Neal McBride, Committee Member, Psychology Faculty
Dr. Melody Moore, University Reviewer, Psychology Faculty

Chief Academic Officer and Provost
Sue Subocz, Ph.D.

Walden University
2020

Abstract

Student Voices in Teacher Evaluation: A Multilevel, Latent Factor Investigation of Teacher Quality

by

Kathleen J. Hoff

MBA, Boise State University, 1998

BS, University of Idaho, 1984

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Psychology

Walden University

May 2020

Abstract

Teacher effectiveness is a key driver of student achievement but persistently difficult to measure. Although precollege student surveys are one cost-effective alternative to traditional observation measures, little empirical research has been conducted on their factor structure. The purpose of this study was to investigate the validity of the Tripod Survey, a widely adopted precollege survey, as a measure of teacher effectiveness. The dataset was the Measures of Effective Teaching (Year 1, N = 1,024 Grade 9 classroom sections, 20,500 students; Year 2, N = 488 Grade 9 classroom sections, 8,658 students). The dynamic model of educational effectiveness guided the study. Multilevel confirmatory and exploratory factor methods were used to evaluate the factor structure of the Tripod Survey at the student- and classroom-levels to determine the degree of construct isomorphism. None of the hypothesized Tripod specifications adequately fit the data at the classroom level and only marginally at the student level. Several alternative bifactor specifications also did not meet minimum requirements for model fit at the classroom level. Additionally, the negatively worded Tripod items appeared to be challenging for students to interpret. These findings suggest that the 36 items composing the Tripod instrument do not capture effective teaching as hypothesized by the Tripod authors. A reduced-form of the Tripod, a two-factor model with 11 items, fit the data well, at both levels. Thus, the full Tripod Survey did not capture a shared perception of students about effective teaching and should be used cautiously to differentiate teachers at the classroom level. This study contributes to positive social change by providing educational leaders with more robust information about the validity of teacher evaluation tools, which can lead to improved learning opportunities for students.

Student Voices in Teacher Evaluation: A Multilevel, Latent Factor Investigation of Teacher Quality

by

Kathleen J. Hoff

MBA, Boise State University, 1998

BS, University of Idaho, 1984

Proposal Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Psychology

Walden University

May 2020

Dedication

To my parents, Hans and Mary Helen, who have always championed my dreams and the efforts needed to see them through.

To my sisters, Helen, Annie, Trish, and Chrissy, who were always there, with a kind word and a glass of wine when I was struggling to remember why I wanted a dissertation.

To my dear friends Mary and Shelly, who were unwavering in their support, incredibly generous with their time, and inspired me to believe I could be a PhD.

Acknowledgments

Completion of this study would not have been possible without the incredible support, encouragement, and generosity of my committee members. First and foremost, I would like to express my deep gratitude to my chair, Dr. David Kriska, whose mentorship and guidance propelled the successful completion of this study and allowed me to become a better researcher and writer.

I would also like to thanks my second committee member, Dr. Neal McBride, for his excellent feedback and support on the various drafts and approval for my study. Thanks also to Dr. Melody Moore, my URR, for her thoughtful and timely reviews.

Finally, I wish to thank my employer and colleagues at the Federal Reserve Bank of San Francisco for their support and encouragement during my course of study.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to the Study

Introduction

Although there have long been concerns about teacher quality, it was not until the No Child Left Behind (NCLB) Act of 2003 that mandatory, nationwide, and systematic evaluations of teacher effectiveness began in earnest (Birman et al., 2009; Huber & Skedsmo, 2016; Wallace, Kelcey, & Ruzek, 2016). A hallmark of NCLB was the shift to a test-based model of school accountability along with a focus on teacher quality as a key input into the education production process (Aldeman, 2017; Hess & Eden, 2017; Schneider, Grogan, & Maier, 2011). The evaluative requirements of NCLB started a new era of high-stakes testing and measurement, making all public school districts accountable for teacher quality and ultimately student learning (Dee & Jacob, 2010; McDonnell, 2015; Taylor, Stecher, O'Day, Naftel, & Le Floch, 2010).

Despite NCLB's intention, its implementation did not always produce the desired outcome (Aldeman, 2017; Dee & Jacob, 2010). Rather than encouraging rigorous evaluation of teacher quality, the NCLB identified "highly-qualified" teachers primarily on educational credentials and the number of years of teaching experience (Birman et al., 2009). With those measures of instructional quality, more than 90% of U.S teachers were considered "highly-qualified" by 2007 (Birman et al., 2009). However, the high level of teaching excellence found throughout the country did not correspond to increasing levels of student achievement as measured by standardized test scores and annual yearly progress benchmarks (Kraft & Gilmour, 2017; Weisberg et al., 2009). The discrepancy between reported levels of teacher quality and student test scores led education officials

at all levels to reevaluate the nature of teacher quality and to seek more effective methods for measuring and evaluating teacher quality (Ravitch, 2013; Rockoff & Speroni, 2010).

The challenge in developing teacher evaluation systems is establishing empirical evidence of instructional effects and transparent procedures for rating individual instructors (de Lima & Silva, 2018). Without evidence that teacher quality can be accurately assessed, the education system will likely not change existing models (Ravitch, 2013; Reddy et al., 2018; Seidel & Shavelson, 2007). Nevertheless, a burgeoning area of K–12 teacher evaluation that has not been widely researched are student surveys designed to assess teaching quality (English, Burniske, Meibaum, & Lachlan-Haché, 2015; Geiger & Amrein-Beardsley, 2019). Advanced as a potential evaluation tool within the federal 2009 Race-to-the-Top grant competitions, precollege student surveys began to gain traction as a viable measure of teacher effectiveness (McGuinn, 2012; U.S. Department of Education, 2009). Combined with traditional measures of teacher effectiveness, student surveys address two key obstacles with more traditional evaluation methods: the high cost associated with direct observation of classroom instruction, typically provided by school personnel or outside evaluators; and the subsequent need to limit the number of teacher observations to contain costs (Balch, 2012; Kuhfeld, 2016; Schulz, Sud, & Crowe, 2014). Limiting the number of observations may introduce bias and reliability issues into important personnel decisions and reduce trust in the evaluation system (English et al., 2015). Further, the inability to cost-effectively evaluate individual teacher competency limits reform efforts, perpetuates a

cycle of low teacher performance, and ultimately reduces lifelong outcomes for students (Chetty, Friedman, & Rockoff, 2014b; Papay, 2012).

Using precollege student surveys to aid in evaluating instructional practices addresses several key challenges in traditional teacher evaluation systems. First, student observations capture hundreds of instructional interactions with a teacher over the course of a school year. Second, the student-based evaluation is provided by a low-cost and captive classroom audience. As such, student surveys offer a potential solution to the time-consuming and costly observation systems traditionally deployed in classrooms to evaluate teacher quality (Marsh, Dicke, & Pfeiffer, 2019). However, the survey instrument needs to deliver valid and reliable evidence of instructional quality. Thus, this study was focused on evaluating the viability of the Tripod Student Perception Survey. The rest of this chapter provides information on the background of the topic, purpose of the study, theoretical framework, nature of the study, and significance of the study.

Background of the Study

The modern era of school reform began just over 50 years ago with the passage of the Elementary and Secondary Education Act of 1965. Signed into law by President Lyndon Johnson, the legislation was a part of the War on Poverty initiative and aimed to reduce inequities in U.S. public education through changes in federal funding and requirements for more equal access to quality education resources (Jennings, 2000; Thomas & Brady, 2005). Since that time, four notable reauthorizations by Congress of Elementary and Secondary Education Act have included (a) the Improving American Schools Act of 1994, which ushered in a new era of standards-based reform efforts

(Thomas & Brady, 2005); (b) the NCLB Act of 2001, which elevated the role of the federal government in establishing requirements and enforcement of accountability (Aldeman, 2017); (c) the Race-to-the-Top grant program of 2009, which incentivized states and individual school districts to implement specific reforms (McGuinn, 2012); and most recently (d) the Every Student Succeeds Act of 2015, which eliminated key NCLB requirements and returned control for rule writing and accountability to the states (Berg-Jacobson, 2016; Pennington & Mead, 2016).

These acts of Congress highlight efforts to reform U.S. public education through federal mandates for equality, standards, accountability, and most recently, a return to local control for decision-making and rule writing (Hess & Eden, 2017; Schneider et al., 2011). Imbedded within each piece of legislation, and of direct importance to this study, are the themes of instructional quality and teacher improvement. Often characterized as a cornerstone of educational reform, teacher quality has been identified as one of the key factors in determining student performance and educational outcomes (Chetty et al., 2014a; Harris & Sass, 2011; Rockoff & Speroni, 2010). Although teacher quality has been at the forefront of educational reform in the United States, establishing systems to evaluate the instructional capabilities of teachers has been difficult and expensive (Weisberg et al., 2009).

Though teacher evaluation has been a challenge, previous legislation has encouraged evaluation measures that may be more effective than what is currently in use. With the 2009 Race-to-the-Top legislation, competitive awarding of grant funds continued to reinforce the focus on teacher quality by incentivizing states and school

districts to develop credible and actionable systems for evaluating their teachers. The formula designed to score Race-to-the-Top grant applications allocated almost one third of the possible points for establishing performance-based measures for evaluating school leaders and teachers. Specifically, performance measures were required to be based on a combination of student achievement scores, observational ratings, and/or student and parent surveys (McGuinn, 2012; U.S. Department of Education, 2009).

By 2015, 43 states had adopted these types of performance-based measures for their teachers (Doherty & Jacobs, 2015). Although performance-based systems had become common, the specifics of individual systems varied greatly from state to state and district to district. For instance, 13 states had mandated specific performance criteria for their evaluation systems, 21 states had required districts to develop a model but left the specific measures to include to local control, and 20 states fell somewhere in between (Hull, 2013). Within these various models, a new but growing practice included parent and student surveys as a component measure of teacher quality, along with more traditional measures that include student achievement scores and classroom observations (Lacireno-Paquet, Morgan, & Mello, 2014). Because most students spend at least 180 days per year in a classroom (Woods, 2015), no other set of observers have more reference points with an individual teacher (English et al., 2015). These approximately 1,100 hours spent with teachers each year represent a potentially valuable source of information about instructional quality. Further, the information and insights collected from student observations represents a relatively inexpensive and captive source of information (Cohen & Goldhaber, 2016; Hinchey, 2010).

Although student and parent surveys of teacher quality may be a viable addition to a teacher evaluation system, there is little empirical evidence on the validity of these types of survey instruments (Geiger & Amrein-Beardsley, 2019; Kuhfeld, 2016; Schulz et al., 2014). This lack of evidence leads to doubt regarding school districts adoption of student and parent surveys within teacher evaluation systems (Spooren, Brockx, & Mortelmans, 2013; Van Der Schaaf, Slof, Boven, & De Jong, 2019). Additionally, without rigorous investigation into the psychometric properties of student surveys, their growing use in teacher evaluation systems remains questionable (Boring, 2017; English et al., 2015; Scherer & Gustafsson, 2015). To address this gap regarding the feasibility of student ratings as measures of teacher quality, I investigated the feasibility of using an existing survey instrument, the Tripod Student Perception Survey (the Tripod), to differentiate teacher instructional quality at the high school level based on an original data set collected by the Bill & Melinda Gates Foundation.

Problem Statement

The need for a high-quality teachers in every classroom has been a cornerstone of educational reforms in the United States for the past 50 years (Chetty et al., 2014a; Harris & Sass, 2011; Rockoff & Speroni, 2010). Prompted in part by comparisons on international student achievement tests, a system that routinely rates over 90% percent of teachers as highly qualified, and recent incentives by the U.S. Department of Education to improve student outcomes, school districts across the country have been embarking on plans to overhaul teacher evaluation systems (Aldeman, 2017; Berg-Jacobson, 2016; Doherty & Jacobs, 2015). A growing measurement component of the new evaluation

systems is the use of student surveys as an indicator of teaching quality (English et al., 2015; Kuhfeld, 2016; van der Steeg & Gerritsen, 2016).

Although precollege parent and student surveys represent a new and relatively unknown class of performance-based measures (Geiger & Amrein-Beardsley, 2019; Herlihy et al., 2014), 31 states have implemented them in teacher evaluation systems, either as a requirement or an optional choice (Geiger & Amrein-Beardsley, 2019). Over a 6-year period from 2013–2019, use of parent and student surveys in the teacher evaluation process grew 258%. Despite several recent studies on the viability of student surveys, primarily at the elementary level, little empirical evidence exists at the secondary level to warrant this rate of adoption by school administrators (Cohen & Goldhaber, 2016; Geiger & Amrein-Beardsley, 2019; Liu, Lindsay, Springer, Wan, & Stuit, 2014). In addition, there is a lack of research into the factor structure of teacher quality at the aggregated level of analysis. Without evidence of a common factor structure between the individual and aggregated levels, it is not possible to directly compare differences among teachers based on the Tripod Survey (Jak, 2016; Jebb, Tay, Ng, & Woo, 2019; Tay, Woo, & Vermunt, 2014). Therefore, I conducted this psychometric study investigating the construct validity of the Tripod student surveys to measure the quality of teacher instruction and effectiveness. This research contributes to the empirical foundation on the use of precollege student surveys to evaluate teaching quality, which can assist education policymakers in making more informed decisions about the appropriateness of including the student's voice in teacher evaluation systems.

Purpose of the Study

The purpose of this quantitative study was to explore the viability of a student survey to serve as a valid and reliable indicator of teacher quality. The results contribute empirical evidence currently unavailable to educational policymakers on student surveys as a component of the teacher evaluation process. My study utilized an existing data set, the Measures of Effective Teaching (MET), collected with funding by the Bill & Melinda Gates Foundation (T. Kane, McCaffrey, Miller, & Staiger, 2013). These data comprise multiple measures of teacher quality, including traditional evaluation tools such as observations ratings and student scores on achievement tests, and a nontraditional approach of incorporating student perceptions of teacher quality at the elementary and secondary levels (White & Rowan, 2014).

I evaluated the construct of teacher quality as measured by the Tripod precollege survey used in the MET study. Developed by Harvard University lecturer R. Ferguson (2012), the Tripod is intended to measure teacher quality, also referred to as teacher effectiveness, across seven dimensions of instructional practices: care, classroom management, clarity, challenge, captivate, confer, and consolidate. I examined the psychometric properties of the Tripod survey (e.g., construct dimensions, levels of analysis, aggregation, and items) in the subjects of English, mathematics, and science at the ninth-grade level. To evaluate the Tripod data, I employed a structural equation modeling (SEM) framework to take into account two levels of data: (a) individual student-level ratings and (b) group-level aggregation of the individual ratings. Using a multilevel confirmatory factor analysis approach, I evaluated the construct validity of the

Tripod to serve as a viable measure of teacher quality (Marsh, Morin, Parker, & Kaur, 2014).

Research Questions

Unlike other multivariate approaches (e.g., ANOVA or multiple regression), SEM relies less on hypothesis testing and more on evaluating theoretical models. I used research questions as the guiding framework for organizing my study, rather than hypotheses, because of the following characteristics of SEM:

- Factor analysis is generally more focused on evaluating the "fit" of specific models as support for a tested factor structure.
- Statistical significance is vulnerable to sample size and, given that SEM is a large sample technique, reliance on the hypothesis testing may be reflective of sample size rather than a meaningful result.
- Calculation of sample and model covariance matrices are completed by computer programs and different programs may use slightly different algorithms. Thus, hypothesis testing around the margins of significance is less meaningful.
- In the social and behavior sciences, there is more interest in the magnitude of a particular effect than in simply testing for significance of the effect. (Kline, 2015).

This quantitative research design was guided by the following research questions:

RQ1: What is the factor structure of the Tripod Student Perception Survey (Tripod) at the student level?

RQ2: What evidence exists to support a higher-order factor structure of teacher effectiveness at the student level?

RQ3: Does teacher effectiveness, as measured by the Tripod, represent a multilevel construct?

RQ4: To what extent does the construct of teacher effectiveness, as measured by the Tripod, exhibit a common factor structure across the student and classroom levels (i.e., psychometric isomorphism)?

RQ5: Does the Tripod exhibit a consistent factor structure across measurement periods?

Theoretical Framework

The two guiding theoretical frameworks used in this study combined distinct but related research in educational effectiveness and psychometric analysis. Educational effectiveness research is found at the intersection of (a) observable teacher behaviors, (b) student learning and achievement, and (c) the situational context of school (Doyle, 1977; Kyriakides, Christoforou, & Charalambous, 2013; Reynolds et al., 2014; Seidel & Shavelson, 2007; Shulman, 1987). Psychometric analysis focuses on the validity and reliability of instruments designed to measure specific constructs (American Educational Research Association, 2014; M. Kane, 1992). Establishing measurement fidelity of the inputs into the educational process more broadly and teacher effectiveness more specifically requires collecting robust, empirical evidence for decisions involving teacher employment and professional improvement. Developing a case for the validity and reliability of the measurement instruments used to evaluate teacher quality are necessary

for differentiating the quality of inputs into the instructional process (Kyriakides et al., 2013).

The dynamic model of educational effectiveness (DMEE) provides a comprehensive theoretical framework for evaluation of the specific factors contributing to the educational process (Creemers & Kyriakides, 2006). The framework captures the complexity of the educational setting by incorporating four key aspects: (a) the primacy of the relationship between teacher and student; (b) the multidimensional nature of instructional practices; (c) the nested, multilevel relationship of students grouped in classrooms, schools, districts, and communities; and (d) the existence of direct and indirect effects of instruction and organization on educational outcomes (Kyriakides & Creemers, 2009). Each of these aspects contributes to student outcomes, so deconstructing their effects is necessary to understanding teacher effectiveness and differentiating instructional quality (Hallinger, Heck, & Murphy, 2014; Weisberg et al., 2009).

Figure 1 displays the theoretical structure of the DMEE. The framework provided an organizing structure to investigate the viability of using the Tripod scores as a measure of teacher effectiveness. The focus of this study is found at the center of Figure 1, labeled "Quality of Teaching." Appropriate measurement of the dimensions of teaching, the measurement model, is an essential component in understanding both the inputs and the outcomes of the instructional process. Without valid measures for the quality of teaching component, this framework cannot be fully realized to evaluate downstream student outcomes.

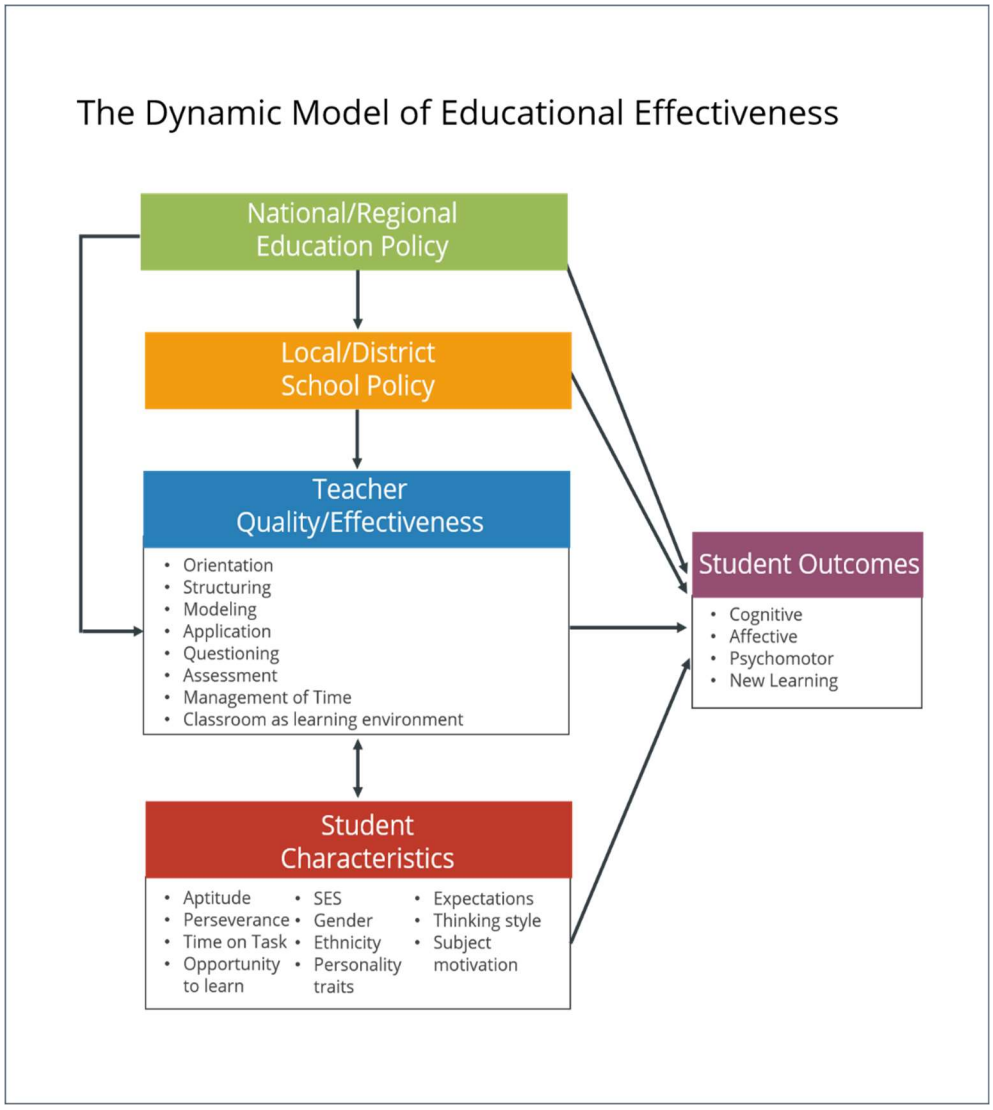

Figure 1. Dynamic model of educational effectiveness.

The DMEE supports this study and its research questions because it situates the evaluation of teacher quality within the multilevel, complex learning environment of schools (Kyriakides & Creemers, 2008; Stapleton, McNeish, & Yang, 2016). Further, this study was focused on evaluating the latent construct of teacher effectiveness. Compounding the evaluation is the manner in which the data are structured, with measurement of the construct occurring at the individual student level and the analysis of the data occurring at the aggregated classroom level. Although this type of multilevel structure is commonplace in educational settings, accounting for that structure within an analytic framework is less common (Zumbo & Forer, 2011).

In addition to the educational effectiveness framework, I used M. Kane's (1992) argument-based approach (ABA) to validation to evaluate the factor structure of the Tripod Student Perception Survey (1992, 2001, 2006, 2013). The ABA involves an iterative, two-step process for establishing validity of a measurement instrument. First, the theory upon which the test is based and its intended use and interpretations are explicitly identified. Second, empirical evidence on the merits of the instruments are then collected and evaluated against the arguments articulated in Step 1 (Hill, Kapitula, & Umland, 2011). As the process continues, evidence accumulates to warrant the use of an instrument in a specific setting (M. Kane, 2016). The ABA aligns closely with the Standards for Educational and Psychological Testing definition of the "degree to which evidence and theory support the interpretation of test scores as entailed by proposed uses of tests" (American Educational Research Association, 2014, p. 69). Thus, M. Kane's ABA strategy for validity provided a well-documented framework to investigate the

psychometric properties of the Tripod Student Perception Survey (Tripod). The research questions correspond to the ABA framework by gathering evidence of the Tripod survey proposed factor structure and mapping the evidence back the survey author's claims (M. Kane, 2013).

Nature of the Study

I designed this quantitative study to examine the feasibility of a precollege student perceptions survey to evaluate teaching effectiveness. My study was focused on student responses to questions regarding the effectiveness of their high school teachers in the subjects of biology, English, and mathematics and used a data set collected by the Bill & Melinda Gates Foundation during the 2009–2011 school year. The variables I used were the item responses from the Tripod collected as student ratings of teacher effectiveness (White & Rowan, 2014).

As will be further described in Chapter 3, I took a multiphased approach to evaluate the Tripod student survey employing confirmatory and, to the extent warranted, exploratory factor analysis to assess the factor structure of the student survey (Kline, 2015). In addition, I employed a multilevel analysis to evaluate the construct of teacher effectiveness at both the student level (Level 1) and classroom levels (Level 2). Finally, I leveraged the two distinct sampling periods, Years 1 and 2, to assess similarities in factor structures across the two periods. Taken together, this quantitative approach based on multilevel factor analysis provided an appropriate methodology to assess the psychometric properties of the Tripod student survey. This analysis offered empirical

evidence not currently available to educational policymakers considering the use of student surveys in the teacher evaluation systems.

Definition of Terms

The following definitions are included to clarify understanding and ensure consistency of meaning of the terms used throughout this study.

Confirmatory factor analysis: An analysis that examines a set of proposed relationships, specified a priori, between observed indicators and their first-order latent variables (Kline, 2015).

Endogenous variable: Dependent variables defined by other variables within a system of equations and denoted by the Greek letter eta (η; Bollen, 1989).

Exogenous variable: Independent variables with no prior causal variables defining them and denoted by the Greek letter *xi* (ξ; Bollen, 1989).

Exploratory factor analysis: In contrast to confirmatory factor analysis, this is an analysis examining potential relationships between observed indicators and latent factors based on empirical modelling (Kline, 2015).

Goodness of fit: A statistical test of how well the observed data fit the hypothesized model (Kline, 2015).

Higher order factor model: Specified when the first-order factors as highly correlated and a generalized, higher-order factor is hypothesized to account for the correlations found in the lower order factors. Also referred to as hierarchical models because the lower-order factors are nested within the higher-order factor (Brown, 2014; Wang & Wang, 2012).

Latent variable: Unobserved variables, constructs, or factors that are defined or measured by a set of observed variables or indicators (Bollen, 1989).

Measurement isomorphism: Indicates that the meaning of a latent construct is similar across levels of analysis and is a necessary requirement for valid comparisons across levels (Ruelens, Meuleman, & Nicaise, 2018).

Measurement invariance: An analytic assessment that a construct of interest is being measured in the same way across groups or measurement occasions. Also referred to as measurement equivalence (Van De Schoot, Schmidt, & De Beuckelaer, 2015).

Measurement model: A specification of the observed variables used to define a latent variable through a regression relationship (Brown, 2014).

Multilevel analysis: Also known as hierarchical or nested data analysis; this type of analysis accounts for distinct parameter estimates at different levels of analysis. An example would be measurement of individual students nested within classrooms, situated with schools. This analysis accounts for measurement error caused by the lack of independence between observed variables (Muthén, 1991).

Observed variable: Also referred to as indicator variables, manifest variables, and reference variables. These items serve to define a latent variable (Kline, 2015).

Psychometrics: A toolkit of statistical methods applied to the construction of assessments that connect an observable phenomenon to a theoretical attributes (Jones & Thissen, 2006).

Reliability: The degree to which an assessment tool produces stable and consistent results (American Educational Research Association, 2014).

Structural equation modeling: A modeling approach that accounts for measurement error in the definition of latent variables obtained through their observed indicators. This approach also provides a framework to investigate relationships between constructs, including direct, indirect, and reciprocal effects (Kline, 2015).

Structural model: A system that represents direct or indirect effects between latent or observed variables and the measurement component of the latent variables with their indicators (Kline, 2015).

Teacher evaluation system: A mechanism to evaluate and administer performance assessments for teaching staff for the purpose of determining employment decisions and professional development needs (Creemers, Kyriakides, & Antoniou, 2013).

Teacher effectiveness: Also commonly referred to as teacher quality, this latent construct refers to the collection of skills and behaviors practiced by an individual educator to produce student outcomes (Creemers et al., 2013).

Validity: The degree to which evidence and theory support the interpretation of test scores for the use of a particular test (American Educational Research Association, 2014).

Assumptions

Implicit in the MET data set are a number of basic assumptions. The first is that the data were collected the manner described in the study documentation. Another important assumption was that high school students had the maturity and thoughtfulness to objectively and to the best of their ability rate the effectiveness of their classroom teacher. Also important was the assumption that teachers were randomly assigned to

classrooms of students in the second year of the MET study. These assumptions represent conventions found in educational settings and are associated with educational research design.

Scope and Delimitations

This study addressed the overarching question of whether high school students are capable of assessing teacher effectiveness through the use of the Tripod (Ferguson & Ramsdell, 2011; White & Rowan, 2014). The Tripod instrument is intended to measure teacher effectiveness across seven dimensions and serve as an appropriate assessment for use in employment decisions and professional development (Ferguson & Danielson, 2014). In addition, the survey suggests that the construct of teacher effectiveness holds across multiple levels of analysis. That is, the meaning of teacher quality as measured by individual student perceptions can be generalized to the classroom level (Lüdtke, Robitzsch, Trautwein, & Kunter, 2009). I investigated these claims through a multilevel, factor analysis that contributed additional empirical validity evidence on the psychometric properties of the Tripod.

The Tripod was employed as one of several assessment instruments in the MET study, conducted at six large, urban U.S. school districts. The school districts involved in the study were recruited via a nationwide request for participation. The sample consisted of non-random intact classes in Year 1 and the random assignment of teachers to classrooms in Year 2. The full MET data set include elementary, middle school, and ninth-grade levels. My study was limited to secondary classrooms represented by ninth-grade biology, English, and mathematics classrooms. The Tripod is based on Ferguson's

(2011) formulation of the factors that make up observable qualities of effective classroom instruction. I empirically evaluated the specific factors of the Tripod and assessed the construct validity of the instrument as a measure of teacher effectiveness at the secondary level. Because of the volunteer nature of the participation by both the school districts and the individual teachers, generalizability may not be fully possible. However, the random assignment of teachers to classrooms in Year 2 of the study provided additional evidence of validity not otherwise available had no random assignment occurred.

Limitations

The MET data set represent a large convenience sample of approximately 3,000 primary and secondary teachers who volunteered to participate in the study as a result of their district's agreement to join the project (T. Kane & Cantrell, 2010). The voluntary nature of the study design naturally limited the degree to which the analysis is generalizable beyond similar settings (e.g., large and urban public schools). Furthermore, the study was restricted to the subject areas of biology, English, and mathematics in Grades 4, 6, 8, and 9, also reducing the ability to generalize beyond similar subjects and grade levels. Further, my research questions were restricted to students in Grade 9.

Another potential limitation is that the source of the dependent variable, student survey responses, were self-reported and may have contained unexamined sources of bias. In addition, the Tripod survey responses were collected in a low-stakes environment, and model performance cannot be generalized to a more high-stake settings environment. Finally, it is important to note that all models are incomplete representations of actual phenomenon (MacCallum, 2001). Teaching is a highly complex,

social interaction that is difficult to model and measure, so all attempts lack precision to some degree.

Significance of the Study

Teacher evaluation has become one of the cornerstones of the educational reform movement, driven by the mounting empirical evidence that teacher quality matters for student achievement (Chetty et al., 2014a; Hanushek, 2011, 2016). Incorporating student perceptions of the learning environment into the evaluation process may enhance discrimination of teacher effectiveness and provide insights for instructional improvement (Goe, Bell, & Little, 2008; Scherer, Nilsen, & Jansen, 2016). The MET data set represented a unique opportunity to evaluate the appropriateness of a high school student surveys as a measure of teacher effectiveness. Further, the evaluations provided by student ratings were potentially a useful tool for informing and improving instructional practices of individual teachers. Adding the student voice to the evaluation process offers valuable and rarely heard insights into what constitutes effective teaching (Chaplin, Gill, Thompkins, & Miller, 2014; English et al., 2015; Martínez, Schweig, & Goldschmidt, 2016; Ravitch, 2013). The first hurdle in establishing such a scale is evaluating the psychometric properties of the survey qualities from a multilevel construct validation perspective (Jebb et al., 2019; Scherer et al., 2016; Tay et al., 2014), which this study addressed.

Summary

In Chapter 1 I introduced the emerging use of precollege student surveys in the evaluation of teacher effectiveness. Traditional measures of teacher quality have

primarily included principal or administrative personal ratings of classroom instruction; however, with the NCLB Act of 2003, issues of in teacher quality arose as a potential area for systematic evaluation and improvement (Birman et al., 2007; Weisberg et al., 2009). Recent iterations in federal legislation have underscored the ongoing efforts to identify high quality teaching. In recognizing that teaching is a complex, multidimensional construct, educational leaders have recognized that no single measure can adequately capture teacher quality (Jensen et al., 2019; T. Kane, Kerr, & Pianta, 2014). Thus, a more comprehensive approach to assessing teacher quality incorporates traditional measures as well as newer methods, including student perceptions of the classroom learning environment.

The brief overview of the Tripod Survey in this chapter introduced the key variable I investigated in this study: the student ratings of teaching effectiveness. Guiding the analysis were two frameworks: (a) the DMEE and (b) the ABA strategy for construct validation. These two frameworks guided the research questions. In addition, this chapter articulated the problem statement, stated research questions, and identified the nature of the study and the assumptions, scope, limitations, and significance for educators, administrators, and policymakers.

In Chapter 2, I review the previous research related to the use of precollege student surveys in the assessment of teacher quality. I begin with an overview of the most recent 50 years of teacher evaluation, followed by a discussion of traditional measures of teacher effectiveness, primarily observation, and rating protocols. The emergence of alternative or nontraditional methods of evaluating teacher performance is examined. The

chapter concludes with the rationale for a study exploring the viability of a student survey as an appropriate measures of teacher effectiveness.

Chapter 2: Literature Review

Introduction

The need for a high-quality teacher in every classroom has been part of educational reforms in the United States for the past 50 years (Chetty et al., 2011; Harris & Saas, 2007; Rockoff, 2004). Although more than 90% of teachers have been consistently ranked as highly qualified nationwide, corresponding student achievement scores continue to vary widely across individual states and schools (Huber & Skedsmo, 2016). In response to the discrepancy between teacher quality ratings and student achievement scores, school district leaders have undertaken large-scale efforts to rethink and rebuild teacher evaluation systems (Harris et al., 2014).

Although commonly used at the postsecondary level, student surveys at the primary and secondary levels represent a new and growing component of teacher evaluation. In 2016, 33 states incorporated student perception surveys within their evaluation systems either as an optional or required component of the teacher evaluation score. Student surveys offer a number of compelling attributes including relatively low costs. Yet little empirical research exists on the appropriateness of this evaluation component (Kuhfeld, 2016; Scherer et al., 2016). The literature on the use of student surveys at the postsecondary level is extensive (Goos & Salomons, 2016; Marsh et al., 2019), but the same cannot be said at the precollege level (Balch, 2012; Kuhfeld, 2017; Wallace et al., 2016). No consensus framework has been developed that explains teacher effectiveness at the precollege levels (i.e., Grades K–12) or the relationship to student achievement (Cohen & Goldhaber, 2016). Though traditional components used to

evaluate teacher effectiveness include principle ratings, classroom observations, and student achievement scores (T. Kane et al., 2014; Muijs et al., 2014), empirical evidence linking student achievement to the construct of teacher effectiveness is lacking in the literature (Hallinger et al., 2014). Within this context, the student point of view, through student surveys, has entered the discussion as a complement to more traditional components of teacher evaluation systems (Martínez et al., 2016; Schlesinger & Jentsch, 2016; Wallace et al., 2016).

The purpose of this literature review is to (a) provide the theoretical and conceptual foundations for a study of high school student surveys to evaluate the effectiveness of classroom teachers and (b) highlight the lack of empirical evidence for the use of student surveys at the secondary level in teacher evaluation systems. In this review, I discuss theories of teacher quality and the lack of empirical research on precollege student perception surveys for teacher evaluation. In addition, I discuss a methodological approach to evaluate the validity arguments for the use of student perception surveys in the assessment of teacher quality.

Literature Search Strategy

I conducted the literature review in two phases: first, a preliminary search designed to yield a broad spectrum of peer reviewed articles for historical context and current practices, and second, a narrower approach focused on the most recent 5 years beginning in 2014 using a more refined set of search terms. Databases included Education Research Complete, ERIC, Sage Premier, and Taylor and Francis Online. In addition, I searched Google Scholar to cross-reference search terms and compare results

to ensure the completeness of the search. I further used the ProQuest Central dissertation database as an additional cross-referencing search tool. All databases were accessed via the Walden University online library.

The literature search included singular and combinations of the following terms: *student survey, student perception, student achievement, teacher evaluation, teacher observation, secondary, high school, teacher effectiveness, teacher evaluation systems, reliability, validity, multilevel analysis, teacher quality, teacher observation, meta-analysis,* and *Tripod Surv*ey. In reviewing the literature, I focused on several concepts identified in my problem statement: teacher evaluation theory, secondary level student surveys, and methodological challenges in measuring latent variables. Much of the pre-2005 research on the use of student surveys in the teacher evaluation focused on analysis at the postsecondary level, where student evaluation of instructors is recognized as an established practice (Marsh et al., 2019). At the precollege levels, growth in the use student perceptions surveys to evaluate teaching effectiveness coincided with the U.S. Department of Education efforts to improve teaching quality (Huber & Skedsmo, 2016; Kuhfeld, 2016; Steinberg & Garrett, 2016).

Theoretical Framework

This study was situated within two fields of research: (a) educational effectiveness and (b) measurement validity. The focus of educational effectiveness research is on the relationship between components of the education process and student learning and achievement (Reynolds et al., 2014; Seidel & Shavelson, 2007). Further, the education effectiveness field provides a framework for examining the impact of teacher

quality in relation to the other components of the educational process (Creemers & Kyriakides, 2006). Within this context, establishing measurement fidelity in educational effectiveness research more broadly and teacher quality more specifically is crucial. In adopting a measure of teacher quality, a case for the validity and reliability of the instrument must be articulated. In this regard, M. Kane's (1992, 2001, 2006, 2013) ABA framework provides a methodology to accumulate empirical validity evidence to support the use of a measurement instrument. Taken together, these two theoretical ideas formed the foundation for investigating the appropriateness of student surveys in the assessment of teacher effectiveness.

Dynamic Model of Educational Effectiveness

The DMEE is a comprehensive theoretical framework used to evaluate factors that impact school-based learning (Creemers, 2002; Creemers et al., 2013). In the absence of theoretical models on educational effectiveness, the DMEE framework was developed to address the need for a theory-based approach to explain the relationship between educational inputs and student outcomes (Vanlaar et al., 2016). In addition, the DMEE incorporates key aspects of the educational environment, including (a) the preeminence of the relationship between teacher and student, (b) multidimensional aspect of instructional practices and characteristics, (c) multilevel analysis indicative of school settings, and (d) the existence of direct and indirect effects of instructional practices on student outcomes (Kyriakides & Creemers, 2009; Reynolds et al., 2014). These four aspects highlight the proximity of the theory itself to the realities of school-based learning.

The theoretical underpinning of the DMEE framework of teacher effectiveness is that teachers are important in determining education outcomes of students (Kyriakides & Creemers, 2009; Vanlaar et al., 2016). Moreover, effective teaching is a complex, latent construct generally identified by specific factors associated with student achievement (Cohen & Goldhaber, 2016; Hallinger et al., 2014; T. Kane et al., 2014). A common conceptualization of teacher quality is as an input into a process that produces student outcomes. As such, some teacher factors are categorized as inputs (e.g., years of teaching experience and content specific knowledge), whereas other factors represent processes (e.g., classroom management and pedagogical skill). The DMEE accounts for this complexity with a latent factor approach reflecting multiple dimensions of teaching quality. In addition, the DMEE recognizes the multilevel organization of schools and accounts for the clustering of students in classrooms, classrooms within schools, and schools within districts, regions, and national settings (Nilsen & Gustafsson, 2016; Scheerens, 2015). Finally, the DMEE acknowledges that direct and indirect effects operate within a multilevel system to influence educational outcomes and the framework seeks to estimate these effects to explicate and improve educational outcomes (Creemers & Kyriakides, 2006).

The viability of a specific student observation instrument to differentiate teacher effectiveness can be measured with the DMEE framework. At the core of the structure are specific factors of teacher effectiveness hypothesized to positively impact student outcomes (Kyriakides et al., 2013). For the DMEE framework, these factors include orientation, structuring, modeling, application, questioning, assessment, management of

time, and classroom as a learning environment. These factors constitute a comprehensive view of teacher behaviors that contribute to student success in the classroom (Creemers et al., 2013). Orientation, structuring, and questioning represents teacher skill in direct instruction, and modeling and application reflect techniques to effectively engage active learning and participation. The classroom learning environment, management of time, and assessment indicate effective classroom environments and the evaluation of student and self (Kyriakides & Creemers, 2009).

Dynamic Model of Educational Effectiveness Related to Current Study

Multiple measures, multiple factors. At the core of the DMEE are the interactions between teachers and students (Creemers & Kyriakides, 2006; Kyriakides & Creemers, 2009; Scheerens, 2015; Vanlaar et al., 2016). Inasmuch as the DMEE is formulated around this critical relationship, accurate measurement of the construct of teacher effectiveness becomes a key component in evaluating differences in student outcomes. Many assessments of teacher effectiveness involve utilizing multiple measures as a best practice (T. Kane et al., 2014). Thus, the DMEE framework is structured to accommodate configurations with multiple measures. In addition, and important to my study, the DMEE posits a multifactor structure for the latent variable of effective teaching (Creemers et al., 2013). The authors of the DMEE argued that their configuration of factors is essentially broad enough to encompass most categories of classroom activities identified as contributing to student outcomes (Kyriakides & Creemers, 2009). This approach provided a framework to investigate the efficacy of the Tripod to evaluate teacher effectiveness.

The DMEE framework has a hypothesized factor structure that includes eight dimensions (i.e., factors): orientation, structuring, questioning, teaching modelling, application, management of time, and assessment (Kyriakides & Creemers, 2008). Six factors relate to instructional practices and two incorporate elements of classroom management and evaluation of students and self (Kyriakides & Creemers, 2009). Similarly, the Tripod hypothesizes a seven-factor structure of effective teaching, including care, confer, captivate, clarify, consolidate, challenge, and classroom management (T. Kane et al., 2014; "Tripod FAQ," 2017). These factors, known as the 7Cs, represent three broader categories of effective teaching and include (a) personal support (factors of care and confer), (b) curricular support (factors of captivate, clarify, and consolidate), and (c) classroom management (factors of academic press, challenge, and classroom management; see Ferguson & Danielson, 2014).

Multilevel application. The DMEE highlights the inherently multilevel aspects of the education process that is both complex and mutable. Important to my study was the linkage between students at the individual level acting as individual raters of their teacher's performance at the classroom level. I differentiated two levels of data: individual student ratings occurring at Level 1, and an aggregation of those ratings (i.e., classroom averages) to represent Level 2. This linkage is defined by student ratings of their classroom teacher, which are aggregated to represent a shared perception of teacher effectiveness at Level 2. This shared perception represents the construct of teacher effectiveness at Level 2 (Fauth et al., 2019; Lüdtke et al., 2009). By design, the aggregated scores from Level 2 become the unit of analysis rather than the individual

student ratings at Level 1. From a psychometric prospective, a key question is whether the within-classroom factor structure for the rating instrument is comparable to the between-classroom factor structure for the instrument (i.e., measurement isomorphism). Without an equivalent factor structures across levels, comparisons at Level 2 become untenable and threaten the validity of the shared construct (Jebb et al., 2019; Stapleton et al., 2016; Tay et al., 2014; van der Scheer, Bijlsma, & Glas, 2019). At issue is the degree to which the meaning of teacher effectiveness can be generalized across levels for valid comparisons between classrooms (Stapleton et al., 2016; van der Scheer et al., 2019).

Aggregating individual data from L1 to measure a group phenomenon at L2 involved a composition model to operationalize the relationship between the two levels (Tay et al., 2014; Van Mierlo et al., 2009). Composition models are deployed in settings when information collected at a lower level is used to make inferences about a construct located at a higher level (Dyer, Hanges, & Hall, 2005). In an influential paper on the topic of composition models, Chan (1998) proposed a five-tiered typology for analysis of multilevel constructs: (a) additive, (b) direct consensus, (c) referent-shift, (d) dispersion, and (e) process composition. The additive model represents a simple summation or averaging of the L1 units, without regard to L1 variation or agreement among L1 units, to form the L2 unit. The direct consensus model represents the L2 unit as a consensus among the L1 units by claiming a functionally isomorphic relationship between levels on the particular construct. Isomorphism is determined by the degree of within-group agreement found at L1 and, once determined, a summation or averaging of L1 items can occur to represent the L2 construct. The referent-shift model is similar to the direct

consensus model in that it relies on agreement among the lower level units. The construct at the higher level, however, is considered to be conceptually distinct from the lower level. In other words, individual level units will assess, rate, or otherwise provide information about a distinctly separate L2 phenomenon (Chan, 1998).

The final two composition models include dispersion and process compositions. Dispersion models, in contrast to consensus and reference-shift consensus models, treat the within-group variance (i.e., the degree of agreement) as a meaningful higher-level construct in and of itself. The within-group variance represents the operationalization of the higher-level construct. The final model within Chan's (1998) typology is the process composition model. Here, the interest is in a particular change that is occurring and identifying a process or mechanism at a lower level that can be also be described at a higher level (Chan, 1998). This composition type is in contrast to the investigation of stable outcomes or attributes found within the previous four models.

For the purpose of my study, the referent-shift model, with students rating their classroom teacher, served as the composition format for investigating the multilevel aspects of the Tripod survey. The reference-shift approach provided a means for assessing the degree to which individual student responses on the Tripod survey can be appropriately averaged to make comparisons between classrooms. The choice of the referent-shift model is consistent with prevailing educational and psychometric research on latent variable measurement in school settings (Dyer et al., 2005; Lüdtke et al., 2009; Stapleton et al., 2016; Zhang et al., 2016).

Although the DMEE is a relatively new theoretical framework, it offers a promising approach to the investigation of effective teaching for its fidelity in representing complexity of teaching within the context of school settings (Nilsen & Gustafsson, 2016). The DMEE framework also provides a broad umbrella under which to evaluate educational effectiveness among the numerous configurations of classrooms and schools that encompass school-based learning.

Argument-Based Approach

In addition to the DMEE, M. Kane's (1992, 2001, 2006, 2013) ABA framework for validation provided an additional lens through which to evaluate the factor structure of the Tripod. The Standards for Educational and Psychological Testing manual defines validity as the "degree to which evidence and theory support the interpretation of test scores as entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests" (American Educational Research Association, 2014, p. 69). M. Kane's ABA was designed to address the key issues in establishing validity through a two-step process addressing the need to accumulate evidence in support of a test instrument constructed along the lines of a specific theory and to articulate the connection between the validity evidence and the test instrument's intended use and interpretation (Chapelle, Enright, & Jamieson, 2010; Cizek, 2016).

The ABA (M. Kane, 1992) grew out of the inconsistent and often confusing definitions of validity and the subsequent difficulty in concluding, empirically, that a given test instrument met the standards of validity. Kane's approach to establishing validity involves a two-part process in which first specific arguments and a rationale for

the use of a given test instrument are explicitly articulated. In the second step, empirical evidence is identified and examined regarding the test instrument. The evidence is then evaluated as in support of or to refute each of the arguments stated in Step 1 (Bell et al., 2012; Hill et al., 2011). This process continues over time to accumulate evidence justifying the instrument as valid within the stated use (M. Kane, 2016).

The Argument-Based Approach Related to Current Studies

Within the context of teacher evaluation, the adoption of the ABA for establishing test instrument validity for traditional measures has been well documented (Cohen, Goldhaber, Grissom, & Youngs, 2016a; Kopriva, Thurlow, Perie, Lazarus, & Clark, 2016). Conversely, the use of any particular approach to documenting the validity of nontraditional student surveys in the teacher evaluation process is much less common (English et al., 2015; Kuhfeld, 2016; Scherer et al., 2016). In general, little empirical evidence exists regarding the validity, reliability, and factor structure of student instruments currently in use to evaluate teacher quality (Geiger & Amrein-Beardsley, 2019; Kuhfeld, 2017; Wallace et al., 2016). This includes the Tripod, one of the most well-known and popular student survey instruments in use in the United States designed to measure teacher quality (Geiger & Amrein-Beardsley, 2019; Kuhfeld, 2016). The MET team included the Tripod survey in its study of teacher quality conducted between 2009–2011 (T. Kane et al., 2013; T. Kane & Staiger, 2012).

Prior to the MET study, virtually no peer-reviewed research had been published regarding validity evidence of the Tripod (Kuhfeld, 2016; Wallace et al., 2016). As it related to my study, the rationale for deployment of the Tripod Survey in more than 30

states is based on little evidence, beyond the reported findings produced by the MET authors, who did not investigate the multilevel factor structure. of the instrument. M. Kane's (2016) ABA provided an established framework from which to examine the instrument's construct validly for evaluating the effectiveness of classroom teachers. The two-part process began by first explicitly identifying the intended use of the Tripod survey (i.e., the validity argument) followed by evidence to support the validity argument. Specifically, the Tripod survey was designed to make inferences about teaching effectiveness that could then be used in teacher evaluation and as a tool for teacher professional development ("Tripod FAQ," 2017). Evidence for the construct validity of the Tripod survey and reliability of student raters had not been clearly established (Geiger & Amrein-Beardsley, 2019). M. Kane's ABA offers a comprehensive and rigorous methodology for building the validity arguments for pre-college student surveys and provide policymakers and school leaders with an evidence based case for the instrument (American Educational Research Association, 2014; Reeves & Marbach-Ad, 2016).

Literature Review Related to Key Variables

Structural Equation Modeling

Teacher effectiveness is by definition a latent construct that can only be indirectly measured. Variables that cannot be directly observed are referred to as latent variables (Kline, 2015) and SEM is a statistical technique that was developed specifically to measure latent variables and to evaluate associations between latent variable and observed variables (Bollen, 1989). As it relates to my study, the latent variable of teacher

effectiveness is hypothesized to be measured by the Tripod student survey with students in the role of observational raters. One of the key assumptions of SEM is that although the latent variables cannot be directly observed, their impact on the other variables can be estimated (Finch & Bolin, 2017). The SEM provides a framework for testing theoretical constructs in the form of confirmatory factor analysis (Kline, 2015). Once an acceptable measurement model is confirmed, subsequent statistical analyses can be conducted (Jak, 2016; Ruelens et al., 2018)

Confirmatory factor analysis is a category of SEM based on evaluating, a priori, an explicit relationship between latent factors and observed indicators (Brown, 2014). The number of factors and pattern of indicator loadings on each factor are specified in advance and is referred to as the measurement model. The model is then evaluated for how closely it reproduces the covariance matrix of the measured variables. The fundamental hypothesis of SEM is that the covariance matrix created by the observed variables is a function of a set of parameters such that, if the model was actually correct, it would exactly reproduce the population covariance matrix (Bollen, 1989).

In confirmatory factor analysis, the model parameters include factor loadings (regression slopes predicting the indicator from the factor), unique variances (variance in the indicator not accounted for by the factor and generally considered measurement error), and factor variances (sample variance of the factor). These basic parameters are used to create the inputs for the variance-covariance matrices describing the associations among the indicator variables and their associated factors. A measurement model is specified by associating specific indicators to specific factors. A model-implied variance-

covariance matrix is then estimated utilizing a "fitting-function," a mathematical procedure to minimize the difference between the sample variance-covariance matrix and the model-implied variance-covariance matrix. Various types of fitting-functions depending on the type of data are represented by the observed variables (continuous, categorical, and binary, as examples). Maximum likelihood is often deployed when fitting continuous data measure models. In the case of categorical data often found in surveys, other estimators are deployed to account for the discrete nature of the rating categories (Muthén, 1991).

Once the measurement model has been estimated (i.e., the model-implied variance-covariance matrix), it is necessary to evaluate fit of the model compared to the sample variance-covariance matrix. Three aspects of the results are used to evaluate the acceptability of the measurement model: (a) the overall goodness of fit, (b) specific areas within the model of poor fit, and (c) interpretability and statistical significance of the model estimates (Brown, 2014). An acceptable measurement model is a prerequisite for any subsequent analysis (Brown, 2014; Kline, 2015).

In the case of the Tripod student survey, I tested the hypothesized measurement model, which includes seven factors as defining teacher effectiveness (Ferguson & Danielson, 2014). In addition, I investigated a series of models implied by the Tripod authors and proposed other researchers (Ferguson & Ramsdell, 2011; Wallace et al., 2016). SEM techniques provided the most appropriate set of tools to investigate the research questions. Further, SEM broadly and confirmatory factor analysis specifically are well documented in the education and psychometric research literature as the key

tools in developing validity arguments for measurement of latent variables (Brown, 2014; Kline, 2015; Nachtigall, Kroehne, Funke, & Steyer, 2003)

Observational Tools for Evaluating Teacher Effectiveness

Traditional observational tools. Teacher evaluation has a long history not only in the United States but worldwide, with an extensive empirical literature (Huber & Skedsmo, 2016; T. Kane et al., 2014; Muijs et al., 2014; Scheerens, 2015). Researchers have identified the classroom teacher as the key level for predicting student achievement (Chetty et al., 2014a; Cohen & Goldhaber, 2016; Hallinger et al., 2014). The challenge, however, has been identifying and measuring differences in teacher effectiveness that can be associated with student outcomes (Cohen, Goldhaber, Grissom, & Youngs, 2016b; T. Kane et al., 2014). The prevailing tool for assessing teacher effectiveness that has dominated education evaluation for decades, is direct observation of teacher practices and subsequent ratings, typically provided by school administrators (Doyle, 1977; Goe et al., 2008; Muijs et al., 2014; Shulman, 1987) As noted in Chapter 1, these traditional rating measures of teacher effectiveness over time have been largely inadequate in differentiating teacher quality and untethered from student achievement scores (Doan, Schweig, & Mihaly, 2019; Jensen et al., 2019; Rockoff & Speroni, 2010).

To address these deficiencies, various models have emerged in the evaluation literature that have attempted to systematically distinguish key components of effective teaching and serve as an observational framework for teacher evaluation (Campbell, 2016; T. Kane et al., 2014; Muijs et al., 2014). Further, these approaches have been designed to be conducted, primarily, by school leaders or other qualified raters

(Campbell, 2016). Four predominate observational protocols currently deployed in school districts throughout the United States and in the MET study (White & Rowan, 2014) are briefly discussed below.

The following models highlight a fairly consistent structure for evaluating effective teaching and are aligned with the dimensions articulated in the DMEE (Kyriakides et al., 2010). While these models have converged around a general set of factors, a number of differences exist in terms of what and how aspects of teacher quality are identified and rated. The most recognized teacher observation model is the Framework for Teaching developed by Danielson in 1996 (Danielson, 2011; Ferguson & Danielson, 2014). The framework is used to deconstruct the complex activity of teaching into components parts that can be observed, measured, evaluated, and ultimately modified for the benefit of students (Danielson, 2011). The theory organizes these components into four domains: planning and preparation, classroom environment, instruction, and professional responsibilities (Danielson, 2011, 2012). The Framework for Teaching provides a general theory of teacher effectiveness suitable for any subject. Other observational protocols also have been developed to measure instructional effectiveness within specific subject areas. These include the Mathematical Quality of Instruction (Heather C. Hill et al., 2008), the Protocol for Language Arts Teaching Observation (Grossman, Loeb, Cohen, & Wyckoff, 2013), and the Quality of Science Teaching (White & Rowan, 2014). Similar to Danielson's Framework for Teaching, each of the subject specific observational tools is comprised of similar components hypothesized to represent effective instruction (T. Kane & Staiger, 2012). In addition,

these models also include subject specific components hypothesized to influences student achievement in those subject areas.

In addition to the formal observational ratings outlined above, school systems have a long history of evaluating teachers via a less formal methodology often referred to as a principal rating (Murphy, Hallinger, & Heck, 2013; Van Der Schaaf et al., 2019). These ratings may or may not involve a specific observation of classroom instructors; instead, the rating may result in a more holistic evaluation by a building principle or other administrator. Numerous authors have found this commonplace practice to be ineffective and often not predictive of student outcomes (Grossman et al., 2013; Hallinger et al., 2014; van der Steeg & Gerritsen, 2016). A number of factors have been hypothesized to explain the weak relationship between principal ratings and student achievement, including rater bias, inadequate commitment to the process, and lack of time (Harris et al., 2014).

Alternative observation tool. In recent years, a new approach to observational measures of teacher effectiveness have been introduced into teacher evaluation systems: the rating of teacher effectiveness by students and parents (T. Kane & Staiger, 2012; Liu et al., 2014; Schulz et al., 2014). Of particular interest to my study was the growing use of pre-college student surveys to evaluate teacher effectiveness (Berg-Jacobson, 2016; Fleenor, 2015; Ross & Walsh, 2019). Similar to traditional observation models of teacher effectiveness, student surveys have been designed to capture key aspects of the instructional effectiveness and are closely aligned with observational models discussed above. The primary difference between traditional observations models and student

surveys is the use of student raters in the role of evaluators. The benefits of allowing students to evaluate teacher effectiveness partly counteract the challenges posed by traditional observation protocols, namely, the high costs associated with one-on-one observations and subsequently limited number of observations that are feasible in a given school year (Cohen & Goldhaber, 2016; Hull, 2013).

One of the most well-known and widely used student survey is the Tripod, developed by R. Ferguson at Harvard University (Geiger & Amrein-Beardsley, 2019). The Tripod survey comprises seven factors hypothesized to measure components of effective instruction (Ferguson, 2010). Organized around three key conceptual categories of personal support, curricular support, and academic press, the Tripod was designed to represent the essential components of instructional practice ("Tripod FAQ," 2017). Each conceptual category is defined by a unique combination of the seven dimensions of teaching practice that comprise the Tripod survey. In addition, each of the 36 items contained in the survey map exclusively to one of the 7Cs: care, confer, captivate, clarify, consolidate, challenge, and classroom management. In the following paragraphs, I briefly describe the three Tripod categories and the corresponding factor or dimension that defines it.

Personal support. Personal support represents the relationship between the teacher and his or her students and the degree to which the classroom environment gives students a sense of being valued and welcomed. This conceptual category is defined by two factors: *care*, that is, showing concerning for students' emotional and academic well-being; and *confer,* encouraging and valuing students' ideas and views. Each factor is

mapped to three and five specific questions, respectively. The message conveyed to students when the care factor is exhibited by the teacher is, "Your success and well-being matter to me in a serious way." The corresponding message with confer is, "I invite, welcome, and respect your ideas and feedback."

Curricular support. Curricular support manifests in teaching strategies that make the curriculum engaging, accessible, and coherent. This conceptual category is defined by three factors: to *captivate*, that is, to spark and sustain student interest in learning; to clarify, that is, helping students understand context and resolve confusion; and to consolidate, or helping students integrate and synthesize key concepts and ideas. Each factor is mapped by four, five, and four items, respectively. The message conveyed to students when the captivate factor exists within the classroom is, "I make lessons intellectually stimulating and relevant." The corresponding message when clarify exists is, "I have multiple good explanations; when you are confused, I will work to help you understand." And when consolidate is practiced, the message is, "I review and summarize lessons to help make learning coherent and memorable."

Academic press. Academic press is evident when classroom conditions foster students' staying focused on achieving their potential. This conceptual category is defined by two factors: challenging, or insisting students persevere and do their best work; and classroom management, or fostering classroom behavior that is orderly, respectful and on-task. Each factor is mapped by eight and seven items, respectively. The message conveyed to students when the challenge factor exists within the classroom is, "I insist upon real understanding, not just memorization, and I will not let you give up, even

when the work becomes difficult." The corresponding message when classroom management exists is, "Our classroom is a place to learn and grow, and I will ensure everyone has an opportunity to focus and concentrate." The conceptual framework for the Tripod student survey is displayed in Figure 2 and discussed below.

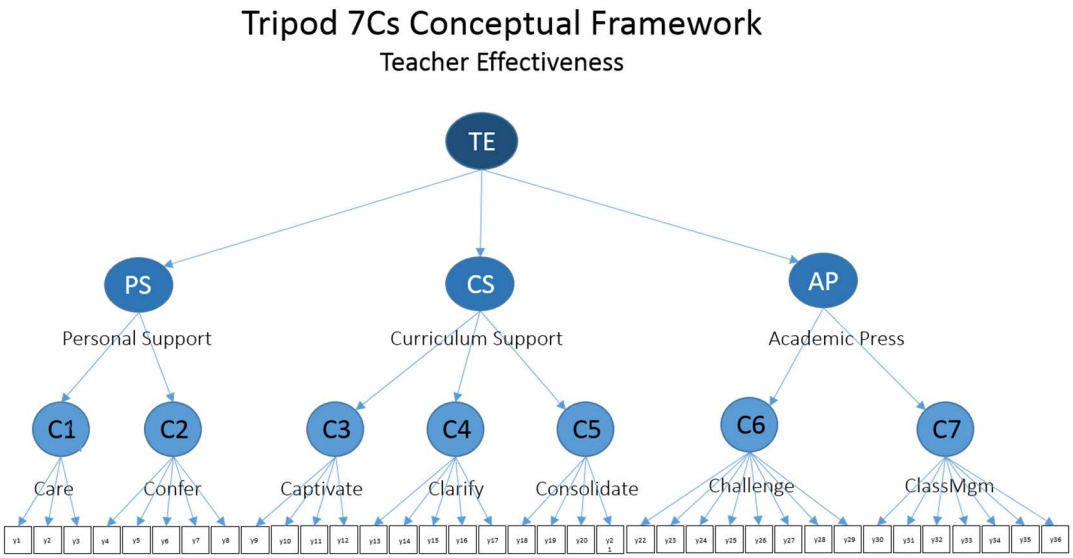

Figure 2. Tripod conceptual framework.

The Tripod Student Perception Survey Framework

The ovals in Figure 2 indicate a higher-order factor structure for the constructs defined within the Tripod survey framework. The top oval represents the general construct of teacher effectiveness defined by the three conceptual categories of the Tripod survey: personal support, curricular support, and academic press. These ovals represent the 7Cs of the framework: care, confer, captivate, clarify, consolidate, challenge, and classroom management. The 36 small, square boxes indicate each of the items that the Tripod student survey comprises. The single-headed arrows from ovals to ovals indicate a direct relationship between a higher level and lower level factor(s). As an example, the

quality of personal support is manifested in the dimensions of care and confer. Further, single-headed arrows from ovals to squares represent which specific dimension or factor of teacher effectiveness is manifested in an observable (i.e., ratable by students) teacher behavior. In this way, student ratings represent the degree to which a particular classroom behavior demonstrated by the teacher then maps back to one of the three conceptual components of teacher effectiveness. An example would be the three items, y1–y3, map to the teaching behavior of "care" that is reflected in the ratings provided by students.

Since its inception in 2001, several hundred thousand K–12 teachers have been evaluated using the Tripod (Tripod Education Partners, 2016). The American Institute for Research (2016) reported that between 2012 and 2015, 4.5 million students evaluated 226,000 classrooms in 29 states. Although the survey has been widely adopted, peer-reviewed research has not been undertaken on the psychometric properties, including its factor structure, scale reliability, and construct validity (Geiger et all., 2019). The MET study, conducted in 2009–2011, was the first documented results of scale reliability. Several dimensions were correlated more to student achievement than other performance measures included in the study (T. Kane & Cantrell, 2010; Kuhfeld, 2016; Scherer et al., 2016). T. Kane and his co-authors (2013) found evidence that the measures of teacher effectiveness, including classroom observations, principal ratings, and student surveys (i.e., Tripod), were correlated with student scores on achievement tests These results have been widely disseminated through various reports available on the Bill & Melinda Gates Foundation website (T. Kane et al., 2013, 2014; T. Kane & Staiger, 2012).

Of interest to my study, the researchers reported that student perceptions, as measured by the Tripod, demonstrated consistency across sections taught by the same teacher, correlations with student achievement scores, and stronger predictions of teacher value-added scores than years of teaching experience or graduate degrees (T. Kane & Cantrell, 2010). These results provide the earliest evidence of the psychometric properties of the Tripod survey for the measurement of teacher effectiveness. These initial findings were encouraging, but additional empirical work is needed to substantiate and further accumulate validity evidence for the Tripod survey. In particular, the lack of a multilevel analysis of the survey factor structure leaves unanswered the fundamental question of what exactly the Tripod is measuring.

Validity evidence for the tripod student survey. The MET study incorporated the Tripod survey as an alternative measure of teacher effectiveness (T. Kane & Cantrell, 2010; White & Rowan, 2014). The findings from the study were published in a series of nonpeer-reviewed reports by the Bill & Melinda Gates Foundation. A database of the study variables was made available to researchers to encourage continued investigations into teacher effectiveness ("Measures of Effective Teaching Longitudinal Database," 2012). Table 1 summarizes the studies on the Tripod for evaluating teacher instructional quality organized by year of publication. The search criteria included the following terms: *Tripod Student Perception Survey, Measures of Effective Teaching, student survey, multilevel analysis, validity, reliability,* and *psychometric propertie*s as well as grade level, methodology, theoretical framework, use of MET data, and significant findings (see Table 1). Nonpeer-reviewed reports and publications were excluded.

Table 1

Previous Peer-Reviewed Studies on the Use of the Tripod Survey

Author (Year)	HS sample?	Theoretical framework	Methodology	Used MET data set?	Key findings
Bradshaw (2017) dissertation	Yes (Grades 6-12)	7Cs (Ferguson)	Descriptive, multiple regression	No	Studied change in teacher scores over a 3-year period on the Tripod survey; the strongest predictor of change was the Y1 score
Wallace, Kelcy, & Ruzak (2016)	No (middle-school math. Grades 6, 7, 8)	7Cs (Ferguson), autonomy support (Reeve & Jang), bi-factor theory	Multilevel, item factor analysis	Yes	Found no evidence of a seven factor model, found evidence of a general factor; suggest Tripod may have limited use in teacher evaluation without additional study
Mabin (2016) dissertation	Yes (high school, Grades 9–11, math)	Student-teacher connection	Descriptive, multiple regression	No	No significant finding in correlations between student-teacher connection and academic achievement
Kuhfeld (2016) dissertation	No (4th-grade math)	7Cs (Ferguson)	Multilevel item factor analysis	Yes	First systematic review of psychometric and validity evidence using a multilevel item factor analysis methodology; findings did not support a 7-factor model
Fleenor (2015) dissertation	No (Grades 3-8 math and reading)	7Cs (Ferguson)	Descriptive; correlational	No	Found statistically significant relationships among two of seven factors ("caring" with reading achievement and "conferring" and math achievement
Polikoff (2015)	Yes (Grades 4-9)	7Cs	Correlation, multiple regression	Yes	Observational measures of instructional quality, including student surveys, exhibited greater stability, year to year, than did value added measures of teacher quality
Schweig (2014b)	Yes (K-12)	Generalizability Theory	Multiple regression	No	Error variance in Tripod survey results are impacted by class and school size

As indicated in Table 1, seven studies, including four dissertations, have been published since the MET study investigating various aspects of the Tripod Student Perception survey. In four those seven studies, researchers analyzed data at the elementary- and middle-school level and three studies used the data from the MET study. Only one study was identified that examined the high-school level (Grade 9) using the MET data. Of the studies summarized in Table 1, Mabin (2016) and Schweig (2014b) examined high-school data to explore group differences, but they did not use the MET data set. Further, Mabin (2016) used a manifest variable approach to test hypotheses about the relationship between teacher caring, as measured by student perceptions, and academic achievement. Only Kuhfeld (2016) and Wallace, Kelcy, and Ruzak (2016) accounted for the multilevel structure of the data, and neither investigated the high-school data. Most important, none of the researchers investigated the necessary equality of factor structures between levels.

Currently, the research literature on the viability of student perceptions of teacher quality is in a nascent stage, with few published studies informing the widespread use of the Tripod in the teacher evaluation process (Geiger et al., 2019). Although the studies provided in Table 1 represent a growing research effort, little work to date has been done to investigate student perceptions at the high school level (Marsh, 2019). Of the seven studies, only Mabin (2016) investigated group differences using the Tripod with high-school students and their teachers. However, that study was limited to a single dimension of teacher effectiveness, the Care factor of the Tripod survey. Moreover, the data represented a convenience sample without random assignment. Furthermore, the

methodology did not account for the multilevel nature of students nested in classrooms nor the latent variable aspect of teacher effectiveness.

My study differed from the studies in Table 1 in three ways. First, I examined high-school students' perceptions of teacher quality with the Tripod using the MET data set. Second, I conducted a multilevel, latent variable analysis to more accurately reflect the nested classroom environment and the indirectly observed construct of teacher effectiveness. And third, I exploited the random assignment of students in Year 2 of the MET study to replicate findings from Year 1.

Summary

In this chapter I presented a review of current literature related to the use of student surveys in the evaluation of teacher effectiveness. I examined the DMEE's theoretical underpinnings and empirical evidence for its use in my study. M. Kane's (1996) ABA to test validity was also examined as a framework to evaluate the psychometric properties of the Tripod. In addition, I reviewed key variables within the MET study. Predictor variables included teacher observation scores, principal ratings, and the Tripod. Criterion or outcome variables included student test scores and teacher value added scores.

Although the Tripod is considered the most well-known and widely deployed pre-college student survey in the United States (Geiger et al., 2019), virtually no peer-reviewed analysis has been conducted on the appropriateness of the instrument for evaluating teacher effectiveness (Kuhfeld, 2016; Wallace et al., 2016). The existing literature includes seven peer-reviewed articles focused on the use of the Tripod. Of the

seven identified studies, three focused on the elementary- and middle-school levels, three were used in K–12 classrooms, and a single study focused exclusively at the high-school level. In a 2016 dissertation, Mabin used the Tripod to study one of the seven factors, Care, to explore group differences among high school students and their teachers. The study yielded mixed results, with no significant relationship between student achievement and the Care factor of the Tripod, and a significant relationship between the Care factor and the race/ethnicity of the student and teacher (Mabin, 2016). Analysis by Schweig (2014b), Flenor (2015), Kuhfeld (2016), and Wallace et al. (2016) also had mixed results and found no evidence to support the 7-factor model proposed by the Tripod authors. Further, Schweig (2014b) found inconsistencies across theoretical models using the Tripod and the importance of justifying the reliability of teacher effectiveness scores. Of note, none of these authors compared factor structure between levels and similarly to the MET researchers, and it appears that equivalent structures were implicitly assumed to exist.

While the desire to evaluate teacher effectiveness has a long history, recent events suggest that teacher evaluation continues to evolve. Developing and implementing evaluation systems capable of capturing the complexity teaching has proven challenging (Huber & Skedsmo, 2016; Ravitch, 2013), highlighting the need to incorporate multiple measures of teacher effectiveness (Hallinger et al., 2014; T. Kane et al., 2014). Within that context, a relatively new measure of teacher effectiveness, pre-college student perceptions surveys, has garnered widespread interest of educational policymakers (Cavanagh, 2014; Geiger & Amrein-Beardsley, 2019; Jensen et al., 2019; Nilsen &

Gustafsson, 2016). Developing a deeper understanding of what student surveys convey about teaching makes an important contribution to the literature on teacher evaluation. In Chapter 3, I discuss the research design and approach, including the use of the MET secondary data set, the data collection procedure, instrumentation, and data analysis. Threats to validity and ethical concerns are also discussed.

Chapter 3: Research Method

Introduction

Teacher evaluation has become one of the cornerstones of the educational reform movement, driven in large measure by the mounting empirical evidence that teacher quality matters for student achievement (Chetty et al., 2014; Hanushek, 2011). Incorporating student perceptions of the learning environment into the evaluation process may enhance the discrimination of teacher quality and provide insights for instructional improvement (Goe, Bell, & Little, 2008; Scherer et al., 2016). Further, adding the student voice into the evaluation process could provide valuable and largely unheard insight into what constitutes effective teaching (Chaplin et al., 2014; English et al., 2015; Martínez et al., 2016; Ravitch, 2013). Thus, the purpose of this quantitative study was to explore the viability of student surveys to serve as a valid and reliable indicator of teacher quality.

This study focused on measurement of the latent construct of teacher quality through use of the Tripod survey, used in the MET study to assess student perceptions of teacher quality. Developed by Ferguson (2012), the Tripod survey is hypothesized to measure teacher quality across seven dimensions of instructional practice: caring, classroom management, clarity, challenge, captivate, confer, and consolidate (Ferguson, 2012).

I explored the psychometric properties of Tripod survey (e.g., dimensions, scales, and items) at two levels of the school environment through a multilevel, factor analytic framework: ninth-grade student responses at L1 and aggregated ninth-grade student responses at L2 regarding algebra, biology, and language arts. Because of the hierarchical

arrangement found in educational settings, with students clustered in classrooms, a multilevel analysis is warranted to account for nonindependence among the student items responses (Lüdtke et al., 2009; Muthén, 1991; Stapleton et al., 2016; Zumbo & Forer, 2011). The results contribute empirical evidence currently unavailable to educational policymakers evaluating the use of student surveys as a component of the teacher evaluation systems. This chapter includes details on my study research design and rationale, the study setting, and sample design for the MET data set. I also describe the Tripod Survey instrument, the data analysis plan, threats to validity, and ethical considerations.

Research Design and Rationale

I used a multilevel factor analysis to investigate the factor structure of the Tripod Student Perception Survey at two levels and to test the construct validity of the survey as a measure of teacher quality. The MET data include multiple measures of teacher quality collected Grades 4 through 9 (T. Kane & Cantrell, 2010). These data were collected over a 2-school-year period from 2009–2011 by the Bill & Melinda Gates Foundation (White & Rowan, 2014). I analyzed the student responses to the Tripod collected at the ninth-grade level from three subject areas: algebra, biology, and language arts. As noted previously, the analysis focused on the degree to which items collected at the student level adequately measure the construct of teacher quality at the classroom level. Multilevel factor analysis and SEM approaches provided broad toolkit of statistical methods to evaluate the relationship among observed and latent variables (Hoyle, 2012; Kline, 2015) and answer the research questions in my study.

The analysis occurred in two phases that leveraged the two sample periods in the MET study. The Phase 1 analysis included the Year 1 sample and addressed RQs 1-4. Phase 2 of the analysis, using the Year 2 sample, addressed RQ5. The data analysis proceeded through a sequence of steps for testing multilevel models and followed a well-documented strategy (Hox, 2010; Muthén, 1991). I also included additional analysis designed to explicitly evaluate configural and metric equivalence across levels (Jak, 2019; Ruelens et al., 2018; Ryu et al., 2009). The strategy generally follows of a series of four steps:

1. Evaluate a confirmatory factor analysis on the total sample covariance matrix (i.e., conventional single level analysis.
2. Determine the need for a multilevel analysis by evaluating the degree of variance at the group level (i.e., the classroom).
3. Estimate separate submodels specified to reflect the same factor structure at the within and between levels of the models, given sufficient group-level variance exists, and to provide evidence of configural equivalence across levels (i.e., configural isomorphism).
4. Estimate a multilevel version of models from the previous step with the added specification of equal factor loadings and provide evidence of metric equivalence across levels (i.e., metric isomorphism).

These steps are designed to be performed sequentially and if at any point in the process the model fails the objective of the given step, the analysis ends with the results from the previous step. For example, if no Level 1 model is confirmed in Step 1, no

further steps are warranted. Similarly, if no variance is detected at Level 1 (Step 2), then the model described in Step 1 is the result of the analysis. As described in Steps 3 and 4, multilevel isomorphism is described by two aspects: (a) equal factor structures and (b) equal factor loadings. These aspects represent configural and metric isomorphism respectively (Jak, 2013; Ryu et al., 2009) and are summarized in Table 2.

Table 2

Measurement Model Isomorphism: Evaluation Criteria

Degree of isomorphism	Number of factors	Pattern of loadings	Rank order of loading magnitude	Item loadings
Weak configural	Equal	Not constrained	Not constrained	Not constrained
Strong configural	Equal	Equal	Not constrained	Not constrained
Weak metric	Equal	Equal	Equal	Not constrained
Strong metric	Equal	Equal	Equal	Equal

Variables

As variables, I used the observed student responses to the Tripod survey, hypothesized to measure aspects of the latent construct of teacher effectiveness. All variables in the data set were linked to specific classroom teachers, providing empirical data about the quality of instruction being provided within individual classrooms, as measured by student perceptions. These rating variables consisted of the item responses to 36 survey questions theorized to measure seven dimensions of classroom instruction. The seven dimensions of classroom instruction include care (three items), confer (five items), captivate (four items), clarify (five items) consolidate (four items), challenge

(eight items), and classroom management (seven items). Each dimension or factor has three to eight indicators reflecting one of the specific subconstructs of teacher effectiveness. (See Figure 3 for the theoretical Tripod framework.) A 5-point Likert response scale was used for all items of the Tripod and the survey was administered across three Grade 9 subjects of biology, English language arts, and mathematics (White, 2014).

Design Choice

Using a multivariate correlational design, I evaluated the theoretical structure underpinning of student perceptions of instructional quality and examined the relationships between individual student perceptions at L1 and the shared perception at the classroom level. Because the Tripod was part of the MET data set, I was able to evaluate the theoretical structure of a relatively untested survey instrument. The use of a multilevel structural equation model approach to validate the hypothesized factor structure of Tripod Student Survey is consistent with the psychometric literature on validity and reliability in educational settings (Kim, Dedrick, Cao, & Ferron, 2016).

A multilevel approach accounted for an important aspect of the research design because measurement at the student level is subsequently evaluated and interpreted at the classroom level (i.e., teacher level). The use of structural equation models in the validation of survey instruments is supported by educational and psychological research (Byrne & van de Vijver, 2014; Morin, Marsh, Nagengast, & Scalas, 2014; Stapleton et al., 2016). A multilevel framework provides an appropriate research design for investigating the context of students nested within classroom (Dunn, Masyn, Johnston, &

Subramanian, 2015; Muthén, 1991). Further, a structural equation analysis represents a theory-driven approach to evaluating relationships among observed and latent variables (Brown, 2014; Heck & Thomas, 2015) and provided a guiding framework to address the research questions in this study (Glaser, 2002; Lam, Schenke, Conley, Ruzek, & Karabenick, 2015; Muthén, 1991; Nachtigall, Kroehne, Funke, & Steyer, 2003).

Methodology

Population

The MET data set was collected from a large sample of teachers in Grades 4–9 during 2009–2011. The teachers were working in six large urban schools districts including Charlotte-Mecklenburg Schools, Dallas Independent School District, Denver Public Schools, Hillsborough Country Public Schools, Memphis City Schools, and New York City Schools. Although the data were not a national random sample, they provide a window into the inner workings of thousands of classrooms across elementary-, middle-, and early high-school settings in a diverse group of urban settings. I examined the Tripod survey of teaching quality across algebra I, biology, and English at the Grade 9 level. Thus, the results appear to be applicable to urban school districts in the United States.

Sampling and Sample

The initial sample for the MET project began with so-called opportunity sampling that occurred between July and November 2009 (White, 2014). The process proceeded in a series of steps beginning with recruitment of school districts, followed by selection of individual schools for participation within each district, and concluded with recruitment of volunteer teachers. Six large, urban school districts agreed to participate in the study.

Within these six districts, schools were identified for participation in the study at the elementary-, middle-, and high-school levels. A number of types of schools within each district were excluded from the study and included special education schools, alternative schools, community schools, autonomous dropout and pregnancy programs, returning education schools, or vocational schools not teaching academic courses. Also exempt from the study were schools that organized teaching in a group or team format, which would make it impossible to connect student learning to a specific teacher. In the final step in the recruitment process, teachers at the targeted grade levels and subject areas were offered the opportunity to volunteer as participants in the study. All volunteers could participate in the study unless (a) they were team teaching, making it impossible to link student learning to the individual teacher; (b) planning to leave the school or switch subjects in the 2010 school year, or (c) fewer than two teachers at the same grade level with the same teaching assignment.

The realized sample for Year 1 (2009–2010 school year) included a total of 2,741 teachers in 317 schools across Grades 4–9 within the six school districts. Of that total, 630 Grade 9 teachers formed the Year 1 subsample for this study. The Year 2 sample (2010–2011 school year) included a total of 2,068 teachers in 310 schools across Grades 4–9. Of that total, 480 Grade 9 teachers formed the Year 2 subsample for this study. Both the Year 1 and Year 2 subsamples had an approximately equal distribution of teachers across the three subject areas: algebra I, biology, and English language arts.

Data Collection

The MET study collected evidence of teaching practices through multiple sources from a volunteer group of teachers in Grades 4–9 in English language arts, mathematics, and science. Evidence of instructional practices were collected at the classroom level from four informant groups including teachers, students, principals, and expert raters trained to evaluate classroom instruction. My study was focused on the student informants at the Grade 9 level, each of whom completed the Tripod for their assigned classroom teacher as part of their end-of-year assessments.

Data Access

For my study, I accessed the MET longitudinal database, a data source located at the University of Michigan that is restricted from general dissemination. The MET data set is secured through the Virtual Data Enclave and required several steps to gain access in addition to conforming to a number of requirements to ensure a secure project office. Steps to gain access the data included (a) submitting an online application via the Inter-University Consortium for Political and Social Research (ICPSR) website, (b) submitting IRB approval/exemption documentation (03-06-18-0151349), (c) scanning and e-mailing a completed confidential data use agreement signed by me and a Walden University institutional representative, and (d) paying a $500 access fee. Further, removing any data analysis or output files was restricted to approval through written requests.

Instrumentation

The MET data set contains multiple sources of ratings of teaching effectiveness based several instruments designed to measure the quality of classroom instruction. The

Tripod represents the variables of interest for this study. Prior to the MET study, no published data regarding the validity or reliability of the Tripod were available. The Tripod website cites findings from the MET study as confirmation of the validity and reliability of the Tripod instrument. My study suggests concerns about the constructs being measured.

Research Questions

RQ1: What is the factor structure of the Tripod Student Perception Survey (Tripod) at the student level?

RQ2: What evidence exists to support a higher-order factor structure of teacher effectiveness, at the student level?

RQ3: Does teacher effectiveness, as measured by the Tripod survey, represent a multilevel construct?

RQ4: To what extent does the construct of teacher effectiveness, as measured by the Tripod survey, exhibit psychometric isomorphism?

RQ5: Does the Tripod survey exhibit a consistent factor structure across measurement periods?

Data Analysis Plan

Once I obtained the data from the ICPSR at the University of Michigan, I screened the data for accuracy and computed maximum and minimum ranges, frequencies, means, normality of distributions, potential outliers, and missing data (Tabachnick & Fidell, 2007). The study data were collected via an opportunity sample from six participating school districts and participant. Following data screening, I

analyzed them using the MPlus 8.0 statistical package and Stata 15.0 statistical package. Prior to conducting any analysis, I implemented a phased approach to evaluate the construct validity of the Tripod (Dyer et al., 2005; Muthén, 1991; Zhang et al., 2016). Appendix A contains specifications and code for the models referenced in the analysis described in the following sections.

Phase 1: Year 1 Sample

Step 1. I assessed the factor structure of the Tripod at L1 only, ignoring the clustering of students within classrooms. This process provided initial information regarding the factor structure of the Tripod survey utilizing the total sample covariance matrix at Level 1 (Muthén, 1991; Ruelens et al., 2018). Using a confirmatory factor analysis, I estimated a series of measurement models. The first model was a base model with a single factor representing teacher effectiveness. The base model assumed all 36 items mapped onto a single latent factor, as shown in Figure 3.

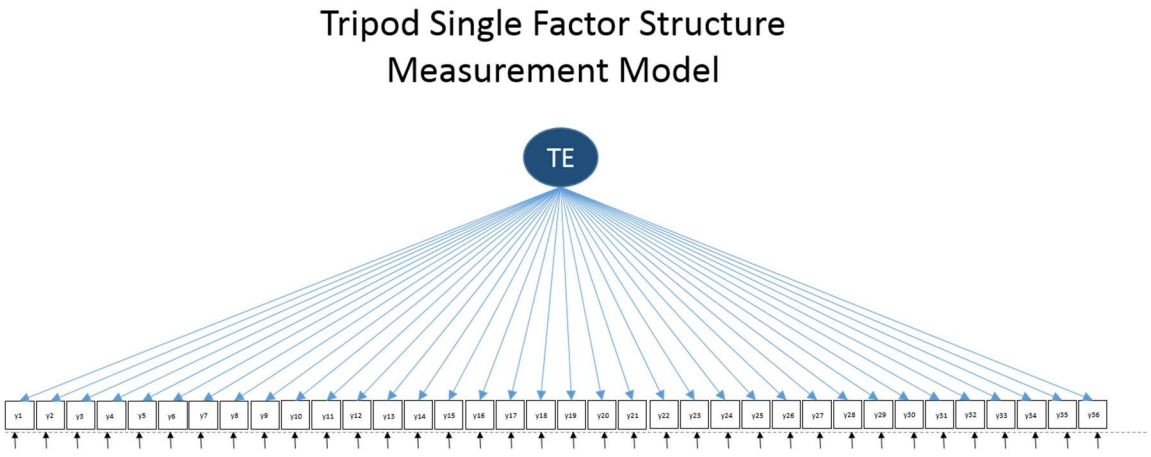

Figure 3. A single factor model.

As shown in Figure 4, the full Tripod model corresponds to a measurement model represented in matrix form as

$$y_i = \tau_i + \Lambda\eta_i + \varepsilon_i \qquad (1)$$

where for individual i, y is a $\rho \times 1$ vector of observed dependent variables, τ is a$\rho \times 1$dimensional parameter vector of measurement intercepts, Λ is a $\rho \times m$ matrix of factor loadings, η is an a$m \times 1$vector of latent variables, and ε is a$\rho \times 1$ vector of measurement errors, uncorrelated with the factors but possibly correlated with other error terms. Further, a covariance matrix of latent factors, ψ and an error covariance matrix, Θ_ε with variances on the main diagonal and covariances on the off-diagonal will be estimated.

Additional models were specified based on the factor structure proposed by the authors of the Tripod survey. Figure 4 represents the seven-factor model hypothesized to map teacher effectiveness via 36 scale items.

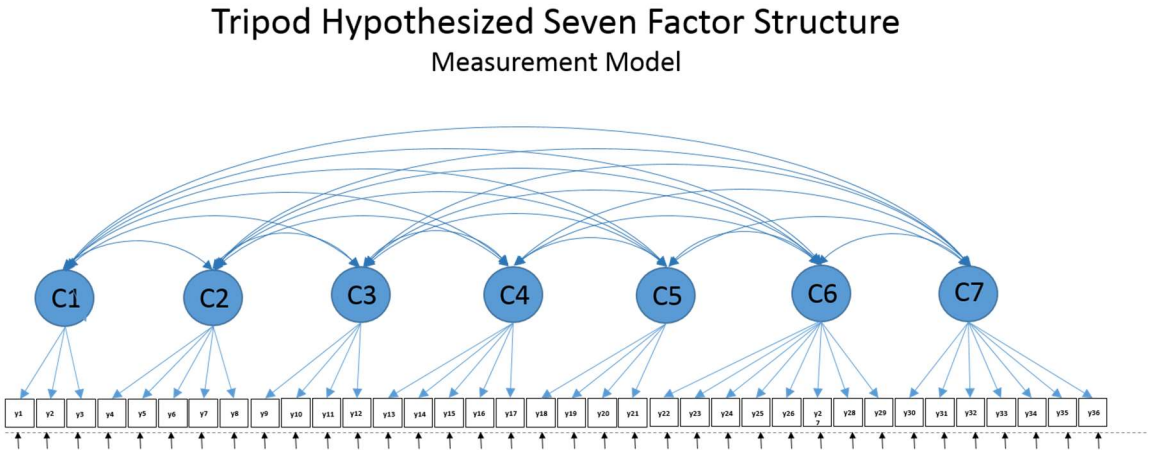

Figure 4. Tripod seven factor model.

The model shown in Figure 4 corresponds to a measurement model represented by seven factors (η_{1-7}) with specific combinations of the 36 Tripod items loading to specific factors. The single-headed arrows originating from each factor (C1– C7) to an observed indicator represent regression paths corresponding to one of 36 equations. An example equation for the path from factor C1 to the observed variable *y*1, is provided below,

$$y1 = \lambda_{11} C1 + \varepsilon_{11} \qquad (2)$$

where λ_{11}, indicates the regression coefficient for the first indicator on the first factor C1, care. The single headed arrows directed toward each indicator represents the residual or measurement error for each equation. The double-headed arrows represent a correlation between two factors. Model 2 specifies the exact pattern of relationships between each factor and indicator, i.e., that is, indicators load onto only one factor, and that all seven factors are correlated with each other. Model 2, in contrast to Model 1, is more restrictive in that it constrains the estimation to reflect the factor structure defined by the Tripod survey author. Further, a comparison of model fit between Models 1 and 2 and additional models will provide empirical evidence regarding the factor structure represented by the MET data.

The confirmatory factor analysis estimated the model-implied covariances for both models, $\widehat{\Sigma_1}$ and $\widehat{\Sigma_2}$, from the observed covariance matrix, S

$$\widehat{\Sigma} = \widehat{\Lambda}\widehat{\psi}\widehat{\Lambda}' + \widehat{\Theta_\varepsilon} \qquad (3)$$

such that the estimated parameters in $\hat{\Sigma}$, were generated by the expected covariances for the observed indicators, $\widehat{\Lambda}\widehat{\psi}\widehat{\Lambda}'$ and the covariance matrix of measurement errors, $\widehat{\Theta_\varepsilon}$

(Bollen, 1989). Models 1 and 2 were evaluated for fit based on traditional fit statistics including chi-squared with degrees of freedom and *p* value; Steiger-Lind root mean squared error of approximation (RMSEA); Bentler comparative fit index (CFI); Tucker and Lewis index and standardized root mean square residual (SRMR; Kline, 2015). These fit statistics showed the relative difference between the data-implied covariance matrix, S, and the model-implied covariance matrix, $\hat{\Sigma}$, for both models (Kline, 2015). Kline (2015) recommended evaluating a suite of fit indices, with suggested cutoff values close to 0.95 or above for CFI and Tucker and Lewis index, 0.08 or below for SRMR, and 0.05 or below for RMSEA, to determine the viability of a given model.

An additional consideration was the 5-point Likert scale in the Tripod survey and subsequent selection of an estimation procedure appropriate for ordinal variables. The framework discussed above applies generally to continuous and multivariate normally distributed variables. Although the data obtained by using a 5-point scale may approach a normal distribution, that assumption may fail to accurately represent the underlying distribution and produce biased parameter estimates (Li, 2016; Liang & Yang, 2014). Under these conditions, normal distribution-based estimation methods are not recommended. That is to say, when maximum likelihood-based estimation methods and Pearson correlations and covariances are conducted on ordinal data, the results may misrepresent the true relationship among variables. The use of categorical threshold models and polychoric correlations can be deployed to account for the non-normal response patterns of ordinal data (Heck & Thomas, 2015; Kline, 2015; Mueller & Hancock, 2008; Rhemtulla, Brosseau-Liard, & Savalei, 2012).

A number of alternative estimation methods are available for categorical ordered variables, including Bayesian with informative priors, Bayesian with non-informative priors, robust maximum likelihood, and weighted least squares with means and variances (WLSMV; Li, 2016; Liang & Yang, 2014). These methods assume that the observed ordinal indicators (y) map to a corresponding continuous latent factor (y^*). This assumption then allows for what are known as threshold models that generate ordinal responses y_i from the latent continuous y_i^* as

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* \leq \tau \\ 2 & \text{if} \quad \tau_1 < y_i^* \leq \tau \\ \vdots & \quad\quad \vdots \\ \text{c - 1} & \text{if} \quad \tau < y_i^* \leq \tau_{c-1} \\ \text{c} & \text{if} \quad \tau_{c-1} < y_i^* \end{cases} \tag{3}$$

where c represents the number of Likert scale categories for y_i, τ_i $(i = 1, 2, \ldots, c-1)$ is the category threshold, and y_i^* is the latent continuous indicator that determines the values of y_i as it crosses different thresholds (Bollen, 1989). Thresholds are estimated through y^* distributions and the sample proportion of responses that fall into each category of y_i. Assuming y^* follows a normal distribution, the cumulative proportions are then converted to values corresponding to a standardized normal distribution and with thresholds. The model relates the ordinal y observations to the measurement model in (1) for y^* with

$$y^* = \tau_i + \Lambda\eta_i + \varepsilon_i \quad (4)$$

where y^* is a $p \times 1$ vector of thresholds, see (1) for comparisons, τ_i represents the threshold structure, Λ is a $p \times m$ matrix of factor loadings, η is an m-dimensional vector of latent variables, and ε is $p \times 1$ vector of measurement errors.

Given the ordinal nature of the observed variables, an estimation method must be selected that can account for the non-continuous nature of the data. As mentioned previously, a number of estimation methods are available. An empirically validated method that has performed well across numerous simulation studies is the WLSMV (Flora & Curran, 2004; Kline, 2015; Li, 2016; Liang & Yang, 2014). The WLSMV estimator fits models to a polychoric correlation matrix and has the advantage of providing model fit indices not available in other categorical estimation methods. While various estimation methods have advantages and disadvantages, the WLSMV has been shown to provide unbiased parameter estimates under a variety of conditions, in particular smaller sample sizes (Li, 2016; Liang & Yang, 2014), and were used to estimate the model parameters in all phases of my data analysis.

Although the Tripod survey authors hypothesized seven factors as defining the construct of teacher quality, little empirical evidence exists to confirm a seven-factor structure (Kuhfeld, 2016; Wallace, Kelcey, & Ruzek, 2016). As a result, I fit a series of models implied by the Tripod authors and suggested by other researchers. Depending on the outcome of the model fit indices, I used aspects of exploratory structural equation modelling (ESEM) to identify an appropriate measurement model for the Tripod, while preserving as much of the hypothesized factor structure as is defensible based the existing

data (Marsh et al., 2009, 2014). I investigated alternative factor structures more representative of the MET data to carry forward into the remaining phases of analysis. In addition, I estimated several higher order models that correspond to the hypothesized second order factor structure of teacher effectiveness. As shown in Figure 5, The second order factors are based on the tripod formulation of three general categories of teacher quality: personal support, curriculum support, and academic press. The 7Cs provide the measurement model estimated in the proceeding step, with the three categories of effective teaching practice represented by second order factors.

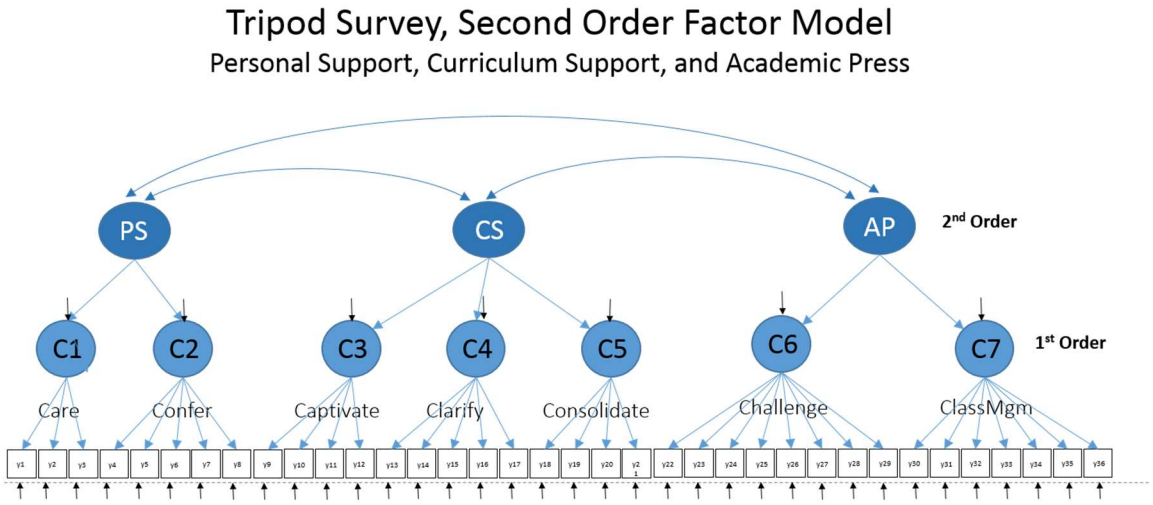

Figure 5. 2nd order 3-factor structural model.

A second possible configuration of a second-order model is based on a single composite score typically generated for individual teachers using the Tripod survey. In this model, a single, second-order factor is assumed to characterize teacher effectiveness, as shown in Figure 6. The difference between the second-order models presented in Figures 5 and 6 is the manner in which the score on individual survey items are combined

to create a composite score. In Figure 6, an individual teacher would receive an average of student responses from a specific grouping of items. As an example, the composite score for Personal Support would be a sum created from the average of the Care items (y1–y3) plus the average of the Confer items (y4–y8). Each teacher would then have a set of three scores generated from specific grouping of the 36 items that make up the Tripod survey. In contrast, the model in Figure 6 generates a single score representing a general factor of teacher effectiveness. The composite score would then be calculated by combining the average score from each of the seven factors. In other words, each factor is equally weighted and contributes equally to the overall score.

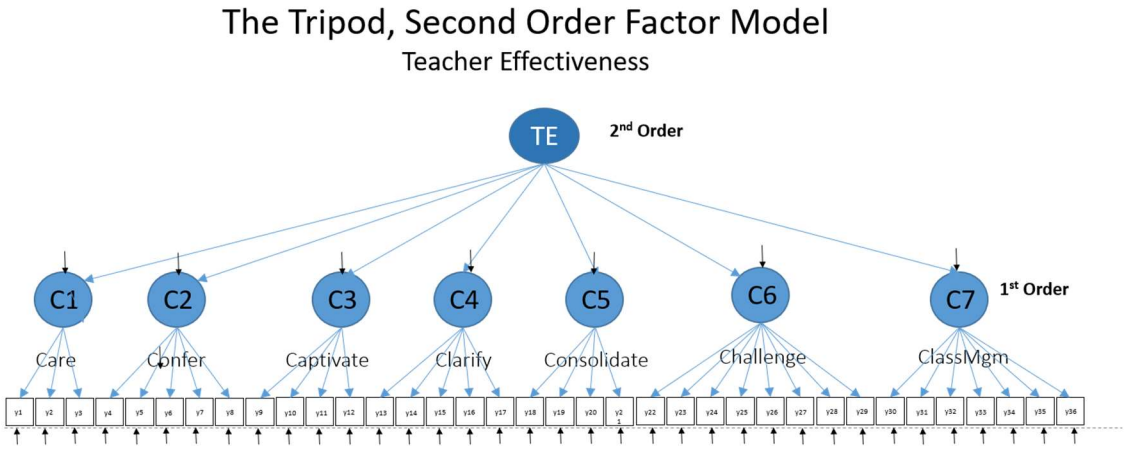

Figure 6. 2nd order, seven factor structural model.

The models shown in Figures 5 and Figure 6 correspond to a structural model represented in matrix form as

$$\eta = \Gamma \xi + \zeta \tag{3}$$

where η is an m-dimensional vector of latent, first order factors, ξ is a $m \times 1$ vector of second-order factors, Γ represents a $m \times n$ matrix of factor loadings of first-order on the

second-order factors, and ζ is an $m \times 1$ vector of latent errors (Bollen, 1989). Note that in Figure 6, all second order factors are correlated as indicated by the double-headed arrows connecting the ovals of personal support, curricular support, and academic press.

Following the work of Wallace, et al., I also estimated several bifactor models, which fall under the higher-order category of models (Wallace, Kelcey, & Ruzek, 2016). The Wallace configuration is shown in Figure 7, highlights a general factor to explain a significant portion of the variance among all 36 observed indicators. In addition to this general factor, Wallace and her coauthors included two additional factors to account for unique variation beyond what is captured by the general factor. These two factors included the factor of control as well as a new factor representing the five negatively worded items in the Tripod. A final point to note is the option of using an item response theory estimation method for the measurement model. Item response theory methodology is well documented in the literature on scale development and validation (Wirth & Edwards, 2007). The focus of the item response theory methodology is at the item level with the aim of investigating the functionality of the scale and less geared toward understanding the structural relationships among factors. Confirmatory factor analysis modeling offers a more flexible framework to evaluate the research questions posed in this study (Wang & Lee, 2016).

Step 2. The aim of this step is to determine the appropriateness of conducting a multilevel analysis by separating the total item variance into the within group and between group variance (Muthén, 1991; Zhang et al., 2016). Three indices are generally examined to determine the degree of L2 variation and provide justification for a

multilevel analysis; interclass correlations (ICC1 and ICC2) and design effect (DE; Heck & Thomas, 2015). The ICC1 can be described as the proportion of total variance of a particular item that can be explained by the between level variation. The ICC (1) is defined as

$$\rho_1 = \sigma_b^2 \,/\, (\sigma_b^2 + \sigma_w^2) \qquad (3)$$

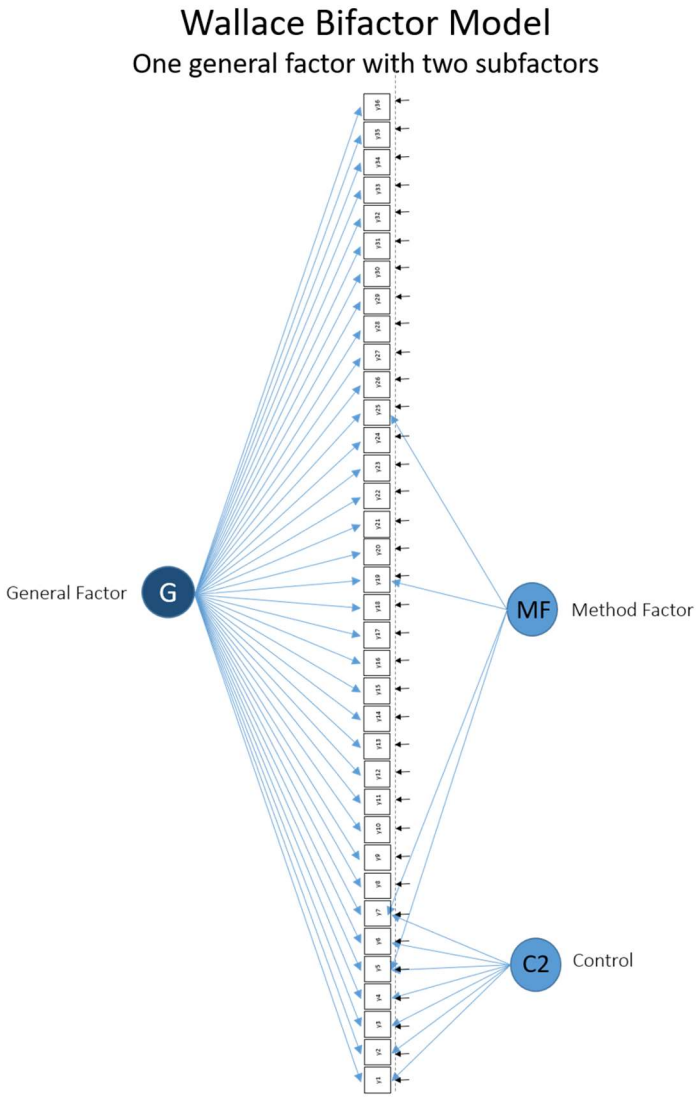

Figure 7. Wallace Bifactor model.

where ρ_1 is the single score intraclass correlation coefficient, σ_b^2 is the between group variance and σ_w^2 is the within group variance. Values of ρ range from 0.0 to 1, with higher values indicating a greater proportion of variance occurring at the between group level and potentially causes bias if the nested structure of the data is ignored. If $\rho < 0.05$, little benefit will be gained from estimating a multilevel model (Dyer et al., 2005). ICC2 is a measure of the reliability of the group means and is defined as

$$\rho_2 = \sigma_b^2 \,/\, (\sigma_b^2 + (\sigma_w^2 \,/\, n)) \qquad (4)$$

where ρ_2 is the average score intraclass correlation coefficient, n is the average group size. Values of ρ_2 range from 0 to 1 with higher values indicating greater reliability. ICC2 values are often evaluated as poor (less than 0.5), moderate (0.5–0.75), good (0.75–0.9, and excellent (greater than 0.95).

A DE, which is a function of average cluster size and the ICC1 and defined as

$$[1 + (ave\ cluster\ size - 1 \times \rho_1)] \qquad (5)$$

where $\rho_1\ is\ ICC1\ and\ is$ calculated for each indicator (item). A DE quantifies the degree to which the sampling error in the clustering of individuals in the study design departs from what would be expected in a simple random sample (Heck & Thomas, 2015). Because clustering exists at L2, individuals are not independent of others within the same cluster. DE greater than 2 have been shown to indicate enough variation at the between group level to support conducting a multilevel analysis. For the purposes of my study, an average $ICC > 0.05$ and $DE \geq 2.0$ across all 36 items for both measures was interpreted as evidence for conducting a multilevel analysis.

Step 3. At this point in the analysis, a multilevel analysis was conducted because I found sufficient evidence of clustering effects in Step 2. The total sample covariance matrix was first decomposed into within and between covariances and separate models were evaluated for L1 and L2. The aim of this step was to determine the degree of configural isomorphism of the factor structure. In other words, how well does the structure at L2 mirror the structure at L1. This analysis provided evidence for configural equivalence which is the first requirement for factor isomorphism. The degree of between- level equivalence of the factor structure was determined by evaluating a series of sub-models for goodness of fit at each level separately. This estimation involves fitting four sub-models, for each given measurement model and calculating level-specific fit statistics, such as RMSEA and CFI. When a given model has acceptable fit at both the within and between levels that provides evidence for identical factor structures across levels (Ryu, 2014; Ryu & West, 2009).

Figure 8 depicts the seven factor Tripod multilevel measurement model with equal factor structures across both levels. Fitting and evaluating the four submodels described above with a seven-factor specification would provide direct evidence of structural equivalence for this hypothesized configuration representing the Tripod measurement model.

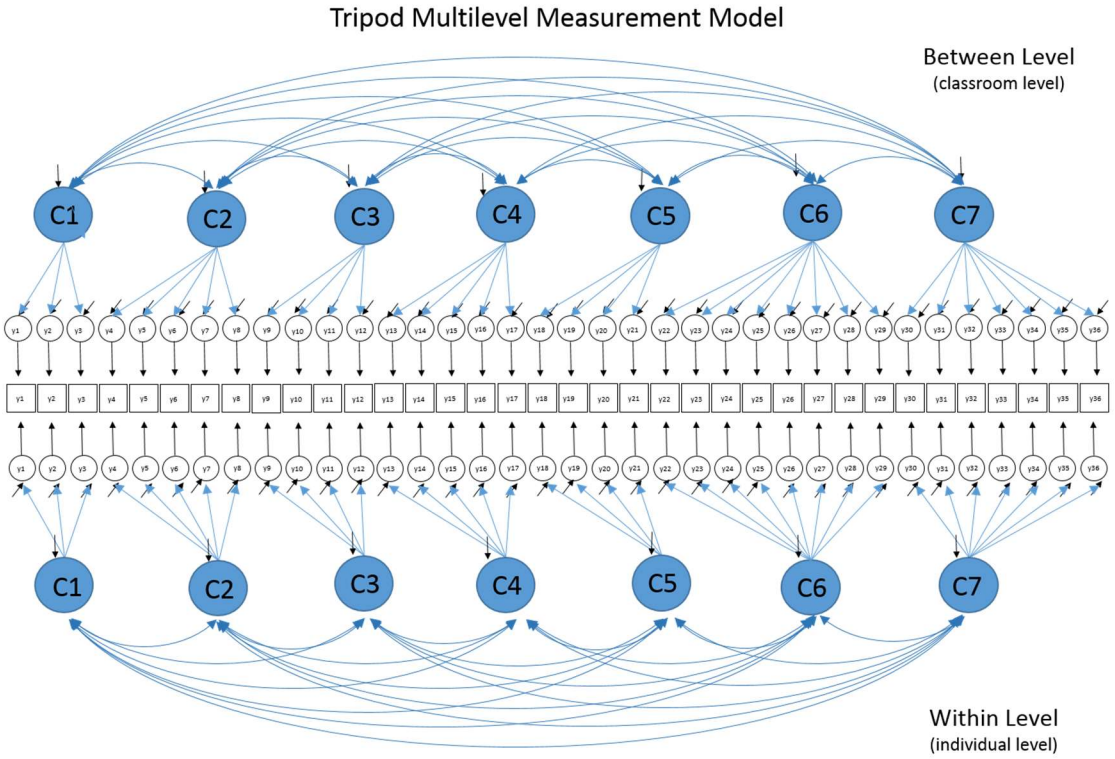

Figure 8. Multilevel Tripod Model.

The multilevel measurement models that were developed are expressed as

$$\gamma_{ij} = \tau_{ij} + \Lambda\eta_{ij} + \varepsilon_{ij} \quad (6)$$

where the multilevel nature of the data is captured by the ij notation, with i signifying the individual student and j used to designate the group membership of the student at the classroom level (Heck & Thomas, 2015). As mentioned earlier, y is the vector of observed items in the Tripod survey, τ is represented by a $p \times 1$ vector of measurement intercepts, Λ is a $p \times m$ matrix of factor loadings, η_{ij} is an m-dimensional set of latent factors, and ε_{ij} is a p-dimensional vector of residuals. The individual, L1, component of the model is not assumed to represent independent observations; however, at the cluster level, L2, independence of the classroom clusters is assumed (Muthén, 1991). The

analysis proceeded by first separating the total sample variance into the within and between variances as

$$\eta_{ij} = \alpha + \eta_{Bj} + \eta_{Wij} \quad (7)$$

where α represents the grand mean for η_{ij}, η_{Bj} represents a random factor component capturing the level two, classroom, effects, and η_{Wij} is a random factor component varying over individuals within their classrooms.

I conducted individual subgroup confirmatory factor analyses, at the within level and between level. These separate estimation procedures generated a pooled sample variance-covariance matrix, S_w for the within level data and a sample between variance-covariance matrix, S_B for the between level data. Model fit statistics were examined to determine which models were configurally equivalent. A final measurement model then was estimated that simultaneously used both the within and between variance-covariance matrices. Factor structure, factor loadings, and errors at both the individual and classroom levels of the data were evaluated. Note that the most viable models, as identified in Step 1, were carried forward into this phase of the analysis.

In addition to the first-order models, I also estimated viable second-order models from Step 1. Figure 9 and Figure 10 are examples of these models.

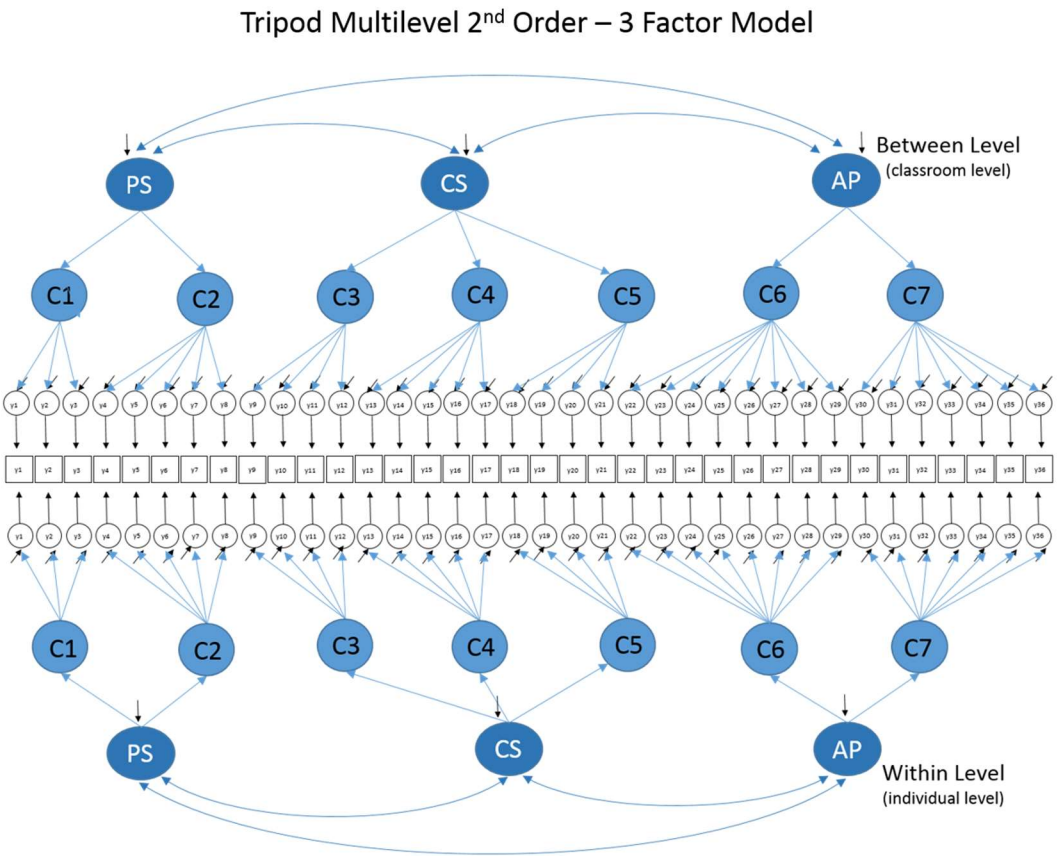

Figure 9. Multilevel second-order model Tripod.

The model shown in Figure 9 corresponds to a second-order multilevel model represented by

$$y_{ij} = \Lambda_W \eta_{Wij} + \varepsilon_{Wij} \tag{8}$$

$$y_j = v_B + \Lambda_B \eta_{Bj} + \varepsilon_{Bj} \tag{9}$$

$$\eta_{Wij} = \Gamma_W \xi_{Wij} + \zeta_{Wij} \tag{10}$$

$$\eta_{Bj} = \alpha + \Gamma_B \xi_{Bj} + \zeta_{Bj} \tag{11}$$

where Equations 8 and 9 represent the measurement model linking observed variables to the underlying factors at each level. Equation 10 and 11 denote the second-order factor at each level. The factor structure, factor loadings, and errors at the second-order can be

evaluated for goodness of fit and compared with the first-order model to determine if evidence exists to support the existence of a high-order factor. In addition and in contrast to the hypothesized three higher order factors, Figure 9 represents a single higher order factor. This model corresponds to the composite score generated by averaging the sub-scores on each of the 7C factors.

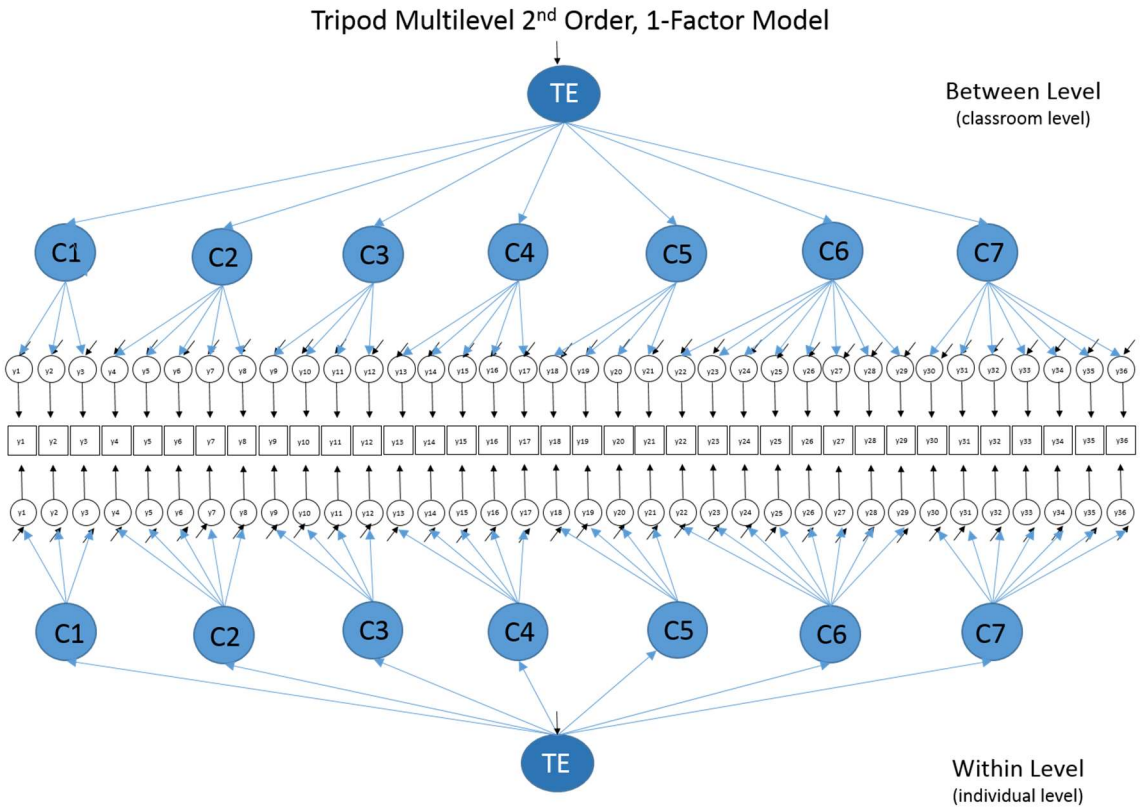

Figure 10. Multilevel second-order, 7-factor model.

Step 4. The final step in this process involved evaluating the degree to which any of the configural models carried forward from Step 3 exhibit metric equivalence across levels. I specified equal factor loading across levels and compared model fit to the model with factor loadings freely estimated. Any decrement in model fit suggests that the factor loadings are not equal across levels and thus not metrically equivalent. Constraining the

model to equal factor loadings across levels represents the concept of measurement equivalence at both levels (i.e., metric isomorphism). Evaluating metric isomorphism provided evidence that the measurement collected at L1 can be aggregated to reflect measurement at L2 for comparisons across groups (i.e., classroom teachers). Without metric equivalence, aggregated student ratings are not valid for comparison at the classroom level (Jak, 2019; Jebb, Tay, Ng, & Woo, 2019; Tay, Woo, & Vermunt, 2014).

Phase 2: Year 2 Sample

In the second phase of the analysis, I evaluated the models developed in Phase 1 using the Year 2 sample. I used the sample methodology described in Phase 1 to evaluate the factor structure of the Tripod survey with the second sample This analysis is designed to contribute additional evidence for the construct validity of the Tripod survey through the use of a second, independent sample (M. Kane, 2016).

Threats to Validity

Several threats to the validity arguments asserted in my study existed. First, the Tripod may fail to capture key components of the latent factor of teacher quality and proposed by the Tripod authors. A second threat to validity of the Tripod was the extent to which factors outside of the intended purpose of the instrument systematically impacted scores (i.e., irrelevant factors; see Reeves & Marbach-Ad, 2016). In addition, some categories of validity evidence cannot be addressed with the MET data set, including item response processes and testing regime consequences. Further, the Tripod responses were collected in a low-stakes environment, and model performance cannot be generalized to a more high-stake testing environment. Although a number of models were

tested and the best fitting model retained, other untested models exist that may have produced similar results. Finally, no covariates were included in the structural model and as such, other influences may well have been operating to influence or bias the observed variables.

Ethical Procedures

All participants in the MET study were recruited through their participating district. Districts were selected based on staff size, central office support, and willingness and capacity to engage in all aspects of the data collection plan, as well as the ability to achieve broad political and union support for the study (Bill & Melinda Gates, 2014). Within the six selected districts, some specific schools were excluded from participation in the study because of highly specialized or targeted student populations or organizational structures that precluded assigning specific teachers to specific students. Teachers in the participating schools volunteered to participate in the study. The MET researchers provided appropriate consent forms and obtained releases from all participating teachers. Parents were informed of the study parameters and provided the opportunity to opt-out of the study (Bill & Melinda Gates, 2014).

The MET longitudinal database requires a signed agreement for the use of confidential data from the MET longitudinal database. The restricted access database is a self-contained computing environment. All data removal and reporting required specific permission. Further, participant data were anonymous, and the signed agreement specified reporting restrictions to protect the individual district information and results.

Summary

This chapter provided research design and methodology for my quasi-experimental, quantitative study on the viability of high school student perceptions, as measured by the Tripod, to provide valid and reliable information about teacher quality. In this chapter I further discussed the research questions and study design. Sampling procedures and selection were outlined, as well as variable instrumentation. I reviewed the two phases of the analysis and methodology for investigating the multilevel measurement models with a brief discussion of model considerations specific to construct validation. Additional information covered several threats to validity including underrepresentation of the latent construct of teacher quality and possibility of irrelevant variance caused by unintended processes. The final section of the chapter covered several ethical considerations stemming from the opportunity sampling procedure and the need to connect individual teachers to specific students. I also discussed the anonymous nature of the data. The findings of the study are discussed in Chapter 4.

Chapter 4: Results

Introduction

The purpose of this quantitative study was to explore the viability of a high school student survey to serve as a valid and reliable indicator of teacher quality. The results contribute evidence not currently available to educational policymakers evaluating the use of high school student perception surveys a component of teacher evaluation systems. I used the MET data set, which was collected during a large study designed to identify multiple measures of teacher effectiveness (Kate et al., 2013). These data included a nontraditional assessment of teacher quality comprising student responses to a 36-item questionnaire assessing the effectiveness of the classroom instruction they received. The research questions were as follows:

RQ1: What is the factor structure of the Tripod at the student level?

RQ2: What evidence exists to support a higher-order factor structure of teacher effectiveness at the student level?

RQ3: Does teacher effectiveness, as measured by the Tripod survey, represent a multilevel construct?

RQ4: To what extent does the construct of teacher effectiveness, as measured by the Tripod survey, exhibit psychometric isomorphism?

RQ5: Does the Tripod survey exhibit a consistent factor structure across measurement periods?

In this chapter, I review the archival MET data set collection methodology as well as the background and demographic characteristics of the participating schools, teachers,

and students in the original study. Descriptive statistics for the participants are followed by summary statistics on the 36-items of the Tripod survey, including means and standard deviations, and frequency distributions, and a discussion of missing data. I evaluate the factor structure of the Tripod survey based on the hypothesized factor structure presented by the survey authors. I also evaluate several higher order models proposed by the survey authors as well as others (Ferguson & Danielson, 2014; Kuhfeld, 2017; Wallace et al., 2016). The results of the analysis are then presented. I conclude the chapter with a summary of the results.

Data Collection and Preparation

The MET study researchers collected data over two school years, 2009–2010 and 2010–2011, from a sample of teachers and students as well as their administrators. The sample consisted of six large urban school districts: Charlotte-Mecklenburg Schools, Dallas Independent School District, Denver Public Schools, Hillsborough Country Public Schools, Memphis City Schools, and New York City Schools. The sampling process for the entire study proceeded in a series of steps beginning with recruitment of school districts, followed by selection of individual schools within each district, and concluded with recruitment of volunteer teachers. The six districts agreed to participate in the study, and within each district, elementary, middle, and high schools, representing Grades 4–9, were identified for participation.

Data Access

The MET data were archived at the ICPSR located at the University of Michigan in 2013. Access to the full, restricted-use longitudinal data was acquired through an

application process that culminated in a data use agreement between ICPSR and Walden University. The agreement outlines strict requirements for access to the database through a confidential online data enclave. No raw data were allowed to be removed from the enclave, and any release or publication of analyses or results required approval by ICPSR administrators. I applied for and received full access to the data in the summer of 2018. All analysis reported here were conducted via the data enclave, and all reporting of results were approved by ICPSR prior to release.

Data Preparation

Within the ICPSR enclave, I was given access to a file repository containing a full set of MET data files. I subsequently copied the files specific to my analysis. An initial review of the data was conducted in Stata, version 15. Because I restricted my study to Grade 9, working databases were constructed to reflect the Grade 9 data and the specific variables outlined in Chapter 3. All descriptive statistics were generated with Stata. For the factor analysis outlined in my five research questions, I converted the Stata files to the specific file format required for execution in Mplus, version 8, via the file transfer program StatsTransfer, version 14.1.

Missing Data

The degree of missing data varied greatly by teacher and student. The number of teachers not reporting characteristics such as gender, years of teaching experience, and advanced degrees varied from a low of 6% missing data on questions of gender and race, up to 58% missing data on years of experience within the district. Missing data for student responses about demographic information varied in a narrow range, by section,

between 11% and 14%. For student responses on the Tripod survey, the amount of missing data varied from item to item with as few as 0.57% to a maximum missingness of 6.1% in Year 1 and from 0.13% to 2.7% in Year 2.

Missing data can be handled in a number of ways during analysis depending on the cause for the missingness. Data are generally classified as missing completely at random, missing at random, and missing not at random. Missing completely at random assumes all missing data are no different than the observed data, and any differences are simple a function of random chance. Missing at random allows for some degree of nonrandom differences between the missing and nonmissing data; however, the cause of missing data is assumed to be unrelated to the variable being measured. Both missing completely at random and missing at random assume, broadly, that missing data can be considered ignorable in relation to potential bias (Kline, 2016). Finally, missing not at random is considered nonignorable because of systematic patterns in missingness that are related to the variable being measured.

I examined missing data patterns using Mplus, v.8.0, which yielded 1,576 distinct patterns in the Year 1 survey responses. Roughly 80% of the 36 Tripod items had a full response set (all questions answered). The most prevalent pattern of missingness occurring 101 times, or 0.49% of total responses, and represented missing responses on three items. The other patterns of missing data occurred less frequently and represented combinations that occurred less than 0.49% within the total responses. In Year 2, missing data exhibited 622 distinct patterns of missingness. Eighty-two percent of the 36 Tripod items had a complete set of responses. The most common pattern of missingness occurred

37 times, or roughly 0.5% of total responses, and represented a missing response on one item. Similar to Year 1, the remaining patterns of missing data represented less than 0.5% of the total responses.

To further explore the potential bias associated with the missing data, I also compared initial model fit indices using pairwise and listwise methods for handling missing data on the individual factors. Pairwise is the default methodology for handling missing data in Mplus when the WLSMV estimator is used. Pairwise deletion uses all available cases in the analysis, whereas listwise deletion deletes all cases with any missing values. I found no discernable difference in model fit using listwise deletion; thus, I concluded that at a minimum the data were missing at random and likely missing completely at random (Brown, 2014; Kline, 2015).

Sample Description

For the purpose of my study, I restricted the analysis to student responses in Grade 9 for a total sample size of $N = 20{,}656$ in Year 1 and 8,122 in Year 2. (See Y2 L2 models.) The descriptive data for individual students were highly similar across both samples, as indicated in Table 3. The students were enrolled in three different courses, Algebra I, biology, and English language arts. The sample size varied by district, with two districts representing roughly 60% of all student responses. Three other districts contributed 40% of the responses, and the final district represented only 11 total responses. Though the individual student responses to the survey items were the key data analyzed in this study, the unit of analysis were classrooms represented by individual teachers. The total number of teachers in Year 1 and Year were 715 and 471,

respectively. Year 1 teachers were distributed across subject areas in rough thirds with slightly less equal distribution in Year 2.

Table 3

Teacher and Section Characteristics

	Year 1			Year 2		
Category	*n*	%	*SD*	*n*	%	*SD*
			Teacher characteristics			
Total number of teachers	715			471		
Male	202	28	16	134	28	2
White	427	60	2	291	62	1.02
Black	142	20	15	95	20	2
Hispanic	42	6	1	28	6	1
Other	35	5	1	22	5	1
Master's degree or higher	433	23	2	112	24	2
Mean years of experience	269	8.8	8.29	187	8.48	8.18
			Section characteristics			
Number of students	1,284	27.7	9.75	480	26.5	7.10
Male		50	0		49	0
White		23	24		22	23
Black		39	30		39	29
Hispanic		29	24		30	23
Asian		6	12		5	10
Other		2	4		2	3
Free and reduced lunch		58	25		58	25
English language learner		9	16		11	18
Special education status		5	9		10	15
Gifted		7	13		6	10

Within the Year 1 sample, 28% identified as male, averaged 8.8 years of teaching experience, and 23% reported holding advanced degrees. The racial composition of Year 1 teachers are as follows 60% White, 20% Black, 6% Hispanic, and 5% reporting as "Other." The Year 2 teachers reported highly similar characteristics to Year 1, which is to be expected given that Year 2 represents the subset of Year 1 teachers who continued with the study into the second year.

The number of sections taught by teachers in Year 1 equaled 1,284 and 480 in Year 2. The average number of students per section was 27.7 in Year 1 and 26.5 in Year 2. The percentage of students receiving free or reduced lunch, by section, was 58% in both Year 1 and Year 2. Nine percent were non-English language learners in Year 1 and 11% in Year 2. Of note, the percentage of students with the status of special education or gifted was 5% and 7% respectively in Year 1 and 10% and 6% Year 2. The racial composition of students, by section, were roughly equally distributed over Years 1 and 2. The mix of gender was also roughly split at 50% over both years of the study.

Tripod Survey Items

The Tripod Student Perception Survey consists of 36 items scored on a 5-point response Likert scale, ranging from *strongly agree* to *strongly disagree*. Five items were negatively worded and were reverse coded to conform to higher scores representing higher degree of agreement with the statement (Kline, 2015). Table 4 displays the survey questions for all 36 items grouped by the hypothesized seven-factor structure.

Table 4

Tripod Student Perception Survey Questions by Factor

Item	Factor	Item ID	Item Wording
1	Care	A10	My teacher in this class makes me feel that s/he really cares about me.
2	Care	B34	My teacher really tries to understand how students feel about things.
3	Care	B146	My teacher seems to know if something is bothering me.
4	Control	B6	Our class stays busy and doesn't waste time.
5	Control	B46	My classmates behave the way my teacher wants them to.
6	Control	B49	Students in this class treat the teacher with respect.
7	Control	B112	Student behavior in this class is under control.
8*	Control	B113*	I hate the way that students behave in this class.
9*	Control	B114*	Student behavior in this class makes the teacher angry.
10*	Control	B138*	Student behavior in this class is a problem.
11	Clarify	B1	If you don't understand something, my teacher explains is another way.
12	Clarify	B17	My teacher has several good ways to explain each topic that we cover in class.
13	Clarify	B80	My teacher explains difficult things clearly.
14	Clarify	B130	My teacher knows when the class understands, and when we do no.
15*	Clarify	B136*	When s/he is teaching us, my teacher thinks we understand even when we don't.
16	Challenge	B21	In this class, my teacher accepts nothing less than our full effort.
17	Challenge	B36	My teacher doesn't let people give up when the work gets hard.
18	Challenge	B45	My teacher wants us to use our thinking skills.
19	Challenge	B59	My teacher wants me to explain my answers.
20	Challenge		In this class we learn almost every day.
21	Challenge	B90	In this class, we learn to correct our mistakes.
22	Challenge	B128	My teacher asks questions to be sure we are following along when s/he is teaching.
23	Challenge	B133	My teacher asks students to explain more about the answers they give.

(table continues)

Item	Factor	Item ID	Item Wording
24	Captivate	B29	My teacher makes learning enjoyable.
25	Captivate	B44	My teacher makes lessons interesting.
26	Captivate	B89	I like the ways we learn in this class.
27*	Captivate	B141*	This class does not keep my attention - I get bored.
28	Confer	A54	My teacher respects my idea and suggestions.
29	Confer	B129	My teacher wants us to share our thoughts.
30	Confer	B135	Students get to decide how activities are done in this class.
31	Confer	B154	My teacher gives us time to explain our ideas.
32	Confer	B155	Students speak up and share their ideas about class work.
33	Consolidate	B58	We get helpful comments to let us know what we did wrong on assignments.
34	Consolidate	B83	The comments that I get on my work in this class help me understand how to improve.
35	Consolidate	B145	My teacher takes the time to summarize what we learn each day.
36	Consolidate	B147	My teacher checks to make sure we understand what s/he is teaching us.

Note. *indicates items that were reverse coded.

Table 5 provides item level descriptive statistics for both Year 1 and Year 2 of the Tripod Student Perception Survey. As mentioned previously, missing data varied by individual item and can be seen in the variation of observations found in Column 2. The number of observations ranged from 20,538 to 19,397 responses in Year 1 and from 8,111 to 7,895 responses per item in Year 2. The mean and standard deviations for Year 1 are provided in Columns 3 and 4. The means ranged from a high 4.2 on item 22 and a low of 2.4 on item 30. The means in Year 2 ranged from 4.27, again on item 22, to the

Table 5

Tripod Item Descriptives

	Year 1			Year 2		
Variable	# Obs.	*M*	*SD*	# Obs.	*M*	*SD*
c1	19,983	3.52	1.21	7,975	3.60	1.20
c2	19,690	3.39	1.17	7,934	3.48	1.15
c3	19,462	2.87	1.26	7,895	2.99	1.28
c4	20,037	3.52	1.13	8,001	3.52	1.13
c5	19,725	3.17	1.18	7,913	3.16	1.16
c6	20,039	3.60	1.13	7,975	3.64	1.14
c7	19,768	3.43	1.22	7,947	3.46	1.21
r_c8	19,710	3.57	1.28	7,945	3.60	1.26
r_c9	19,508	3.18	1.28	7,926	3.14	1.27
r_c10	19,917	3.50	1.26	7,977	3.54	1.21
c11	20,017	3.86	1.08	8,012	3.92	1.04
c12	19,921	3.76	1.07	7,992	3.82	1.06
c13	19,543	3.66	1.10	7,938	3.77	1.07
c14	20,061	3.69	1.08	7,998	3.71	1.06
r_c15	19,781	3.52	1.15	7,940	3.57	1.11
c16	19,397	3.73	1.09	7,900	3.88	1.05
c17	19,958	3.76	1.09	7,985	3.82	1.11
c18	19,737	3.89	1.03	7,937	4.01	0.99
c19	19,909	3.83	1.05	7,978	3.93	1.02
c20	19,673	3.76	1.06	7,938	3.82	1.03
c21	19,481	3.75	1.05	7,906	3.86	1.04
c22	20,406	4.20	0.98	8,050	4.27	0.96
c23	20,101	3.92	0.98	8,013	3.95	0.96
c24	20,014	3.46	1.25	7,990	3.47	1.23
c25	19,756	3.43	1.20	7,959	3.50	1.20
c26	20,538	3.69	1.07	8,111	3.75	1.04
r_c27	19,764	3.22	1.32	7,954	3.31	1.29
c28	19,461	3.67	1.10	7,922	3.78	1.09
c29	20,369	3.79	1.11	8,064	3.92	1.09
c30	20,445	2.42	1.04	8,067	2.45	1.03
c31	19,398	3.56	1.08	7,902	3.69	1.05
c32	19,468	3.57	1.13	7,916	3.67	1.11
c33	19,799	3.51	1.17	7,967	3.64	1.13
c34	19,816	3.52	1.15	7,979	3.59	1.13
c35	19,638	3.44	1.15	7,919	3.46	1.16
c36	19,583	3.84	1.07	7,928	3.94	1.04

Note. The prefix r_ indicates a reverse coded item.

low end of 2.4, this time, on Item 19. The range of ICCs are found in Columns 5 and 9, for Years 1 and 2, respectively. In Year 1, the ICCs fell into a range between .310 to .107.and in Year 2 from 0.344 to 0.143. The items were all scored on a Likert scale from 1–5, with 1 as the lowest and 5 as the highest.

Study Results

Overview of Analysis Plan

The organization of Research Questions 1–4 mirrors a well-documented strategy for the analysis of multilevel models (Dyer, Hanges, & Hall, 2005; Hox, Moerbeek, & Van de Schoot, 2017; Muthén, 1991; Stapleton, McNeish, & Yang, 2016). The strategy consists of four steps:

1. Evaluate a confirmatory factor analysis on the sample total covariance matrix (i.e., conventional single level analysis).
2. Determine the need for a multilevel analysis by evaluating the degree of variance at the group level (e.g., classroom).
3. If sufficient group level variance is found, then estimate separate within and between level submodels to determine the if the factor structures demonstrate configural equality (e.g., Level 1 and Level 2 have identical factor structure).
4. Estimate simultaneous multilevel models to determine metric equality (e.g., corresponding factor loadings across levels are identical).

These steps are performed sequentially, and if at any point in the process the model fails the objective of the given step, the analysis ends with the results from the previous step. For example, if no Level 1 model is confirmed in Step 1, no further steps

are warranted. Similarly, if no variance is detected at Level 1 (Step 2), then the model described in Step 1 is the result of the analysis. As noted in Chapter 3, multilevel isomorphism consists of two primary considerations across the levels: (a) equal factor structure and (b) equal factor loadings. These considerations are termed configural and metric isomorphism respectively and are summarized in Table 6.

Table 6

Criteria for Measurement Model Isomorphism

Degree of isomorphism	Number of factors	Pattern of loadings	Rank order of loading magnitude	Item loadings
Weak configural	Equal	Not constrained	Not constrained	Not constrained
Strong configural	Equal	Equal	Not constrained	Not constrained
Weak metric	Equal	Equal	Equal	Not constrained
Strong metric	Equal	Equal	Equal	Equal

Note. Criteria for evaluating equality across levels.

Data Analysis

In the research questions that follow, I addressed the steps in the multilevel evaluation strategy such that RQ1 and RQ2 addressed Step 1, RQ3 addressed Step 2 and Step 3, and RQ4 addressed Step 4. In RQ5, I tested my findings from Year 1 with the Year 2 sample.

Research Question 1: What is the factor structure of the Tripod Student Perception Survey? To determine the factor structure of the Tripod survey, I tested a series of single-level confirmatory factor models, reported in Table 7. While the multilevel nature of the data is clear, best practices in multilevel SEM is to begin with the simpler, single level data, in order to establish an acceptable measurement model prior to testing the multilevel structure (multiple citations here). Further, each model in Table 7

represents a hypothesized factor structure, proposed either by the Tripod authors (Ferguson & Ramsdell, 2011; Ferguson, 2012) or by others (Wallace et al., 2016). All models were evaluated based on a suite of recommended goodness of fit indices including model chi-square with degrees of freedom and *p* value; root mean square error of approximation (RMSEA) and 90% confidence interval; and the CFI (Brown, 2014; Kline, 2015). Methodologists recommend a range of acceptable fit values when evaluating the viability of a model. Because of known susceptibility to large sample size, I report model chi-square for all models but do not interpret the significant *p* value as an indication of model misfit (Brown, 2014; Kline, 2015). For the other indices, however, I follow generally recommend ranges, including RMSEA close to 0.06 or below as suggesting good model fit, 0.08–0.10 as mediocre, and models with RMSEA greater than or equal to 0.10 are deemed poor fitting and not interpreted (Bollen, 1989; Brown, 2014; Byrne, Shavelson, & Muthén, 1989; Kline, 2015). CFI and TFI close to .95 or above, are interpreted as representing good model fit, .90–.95 as acceptable and values below .90 to be questioned. (Brown, 2014)). Finally, SRMR close to .80 or less are interpreted as indicative of good fit.

As a preliminary step, I first evaluated the seven unidimensional constructs of the Tripod survey: care, control, clarify, challenge, captivate, confer, and consolidate. This analysis of the component factors of the Tripod survey is useful as a first step in understanding the more complex hypothesized models that follow. The confirmatory factor analysis of the seven individual factors yielded mixed results. Two factors, Clarify and Captivate, fell into the ranges of good to excellent fit to the data on all four fit indices

(see Table 7). Two of the constructs, Challenge and Confer, provided reasonably good fit to all indices, with Control and Consolidate demonstrating poor to unacceptable ranges of fit across indices. Note that the Care factor has only three indicates and could not be evaluated for fit. A closer look at the poor fit of the Control and Consolidate factors revealed no issues with localized areas of strain, as the standardized residuals were all less than the absolute value of .10 (Kline, 2015). Although the majority of the Tripod individual factors exhibit mixed to poor fit, I continued with model testing that included the seven factor Tripod model with the caveat that subsequent results should be interpreted with caution.

I began the model analysis with a single-factor model as a starting point for comparison purposes (Brown, 2014; Muthén, 1991), then estimated four subsequent models based on factor structures implied by the Tripod practitioners (White & Rowan, 2014). These models were specified in accordance with common confirmatory factor analysis such that each item loads on only one factor, all cross-loadings are constrained to zero, and factors are allowed to covary (Brown, 2014). Model 1 (Figure 3) represents the Tripod survey as a unidimensional construct of teacher effectiveness with all 36 items loading onto a single factor. This model exhibited poor fit to the data (RMSEA = .087, CFI = .723, SRMR = .069). Model 2 denotes the three factors representing broader teacher qualities of academic press, curriculum support, and personal support and demonstrated a similar lack of fit (RMSEA = .084, CFI = .743, SRMR = .066). Model 3 provides the configuration most generally referred to the Tripod Survey (Figure 4) and posits seven factors to describe effective teaching. This model also did not yield strong fit

to the data (RMSEA = .062, CFI = .863, SRMR = .045). Although neither Model 2 nor Model 3 exhibited consistent levels of fit across all three fit indices, Model 3 did provide some evidence of improvement over the single factor and three factor configurations. That said, the very low values for CFI strongly suggests each of the models is miss-specified.

Given the theoretical importance of the seven-factor model in assessing teacher effectiveness (Geiger & Amrein-Beardsley, 2019) and the ambiguous results, I examined possible sources of misfit within Model 3 rather than outright rejection (Persson, Kajonius, & Garcia, 2019). In the confirmatory factor analysis literature, diagnosing potential areas of misfit in a model solution suggest examining standardized residuals and modifications indices. A review of the residuals for Model 3, which ranged from -.163 to 2.38, yielded no discernable issues and were almost all below the 1.96 level for significance, providing evidence that the indicators were generally reproducing the model parameters (Brown, 2014). The modification indices, on the other hand, suggested 166 possible changes to the specification of the model that would result in a reduction in the model χ^2 estimate from 10 to 7,049 over the initial model. Of the total changes, 35 re-specifications would yield a reduction in model chi-square from 1,000–7,000 and involve cross-loading items between multiple factors. As an example, freeing item C8 to loading on all seven factors is expected to reduce model χ^2 by 41,686.02 units. However, without a substantive theoretical basis for releasing this and the many other parameters suggested by the modification indices, it is considered inadvisable to re-specify the model (Brown, 2014). And to again, take caution in interpreting parameter results.

Based on the degree of possible misspecification in Model 3, I also estimated an ESEM. This exploratory framework has been recommended for confirmatory factor analysis with complex models (many highly correlated factors) that fail to reach acceptable levels of model fit (Asparouhov & Muthén, 2009). Model 4 mirrors the original seven-factor specification, with each of the 36 items loading on a specific factor. Unlike the previous confirmatory factor analysis configurations, with the ESEM specification each item cross-loads onto each nontarget factor. These cross-loadings, however, are targeted to be as near zero as possible through the use of an oblique rotation process (Asparouhov & Muthén, 2009). The seven-factor ESEM model (Model 4) achieved a very good level of fit (RMSEA = .036, CI for $p < .05$ = 1.00, CFI = .987, SRMR = .012), surpassing the fit of Model 3.

The final model estimated consisted of two of the individual factors; Clarify and Captivate (Model 5, see Figure 11). Of the seven factors hypothesized by the Tripod authors, these were the only two factors that exhibited good levels of fit across all fit indices. Although not formulated as a possible factor structure by the Tripod authors, I evaluated this model because on the strength of the two individual factors and the marginal fit of the seven-factor model. The results of Model 5 (two-factor reduced form) provided excellent fit across all three fit indices (RMSEA = .047, CFI = .973, SRMR = .011).

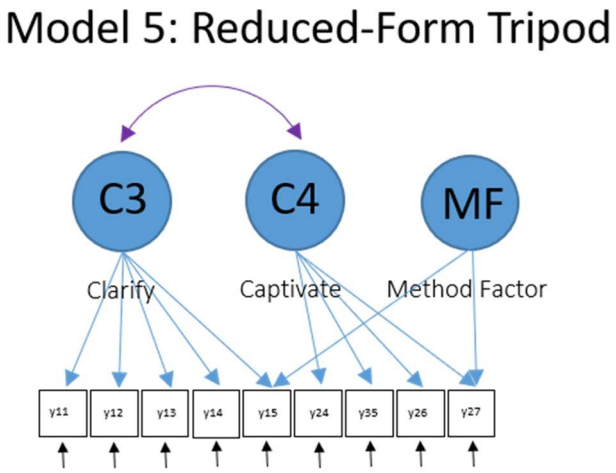

Figure 11. Model 5: Reduced-form Tripod.

To summarize, the purpose of RQ1 was to conduct Step 1 in the multilevel confirmatory factor process by confirming a single level, first order model with acceptable levels of fit. Three models achieved a degree of mixed to good fit including, the seven factor Tripod model (Model 3), its hybrid counterpart the ESEM seven factor model (Model 4), and a two-factor model representing the constructs of Clarify and Captivate (Model 5), based on generally acceptable levels of model fit. See Table 7 for all model fit indices. I proceed in the next section (RQ2) to determine if the Year 1 sample data support a higher-order factor structure at the individual level for the seven-factor Tripod model.

Table 7

Year 1 Test of Model Fit: Single Level

Models		χ^2	*df*	RMSEA (90% CI)	CFit	CFI	SRMR
			Individual Factors			-	- -
Care	C1-C3	53,391.789*	0	-	-	-	-
Control	C4-C10	4,763.607	14	.130 (.127-.133)	0.0	.903	
Clarify	C11-C15	232.834	5	.048 (.043-.053)	.760	.995	
Challenge	C16-C23	2,489.617	20	.078 (.076-.078)	0.0	.954	
Captivate	C24-C27	103.85	2	.050 (.042-.059)	.467	.998	
Confer	C28-C32	563.058	5	.074 (.069-.080)	0.0	.984	
Consolidate	C33-C36	730.315	2	.0135(.027-.135)	0.0	.981	
			Hypothesized Factor Structures				
Model 0	Baseline	191,604.758	630	.268			
Model 1	1 Factor	91,649.871	594	.087 (.087-.087)	0.0	.723	.069
Model 2:	3 Factor	85,001.353	591	.084 (.084-.084)	0.0	.743	.066
Model 3	7 Factor	45,691.744	573	.062 (.062-.063)	0.0	.863	.054
Model 4	7-Factor ESEM	9,627.339	399	.034 (.033-.034)	1.00	.972	.012
Model 5	2-Factor	1,155.072	25	.047 (.045-.050)	.973	.990	.011

Note. N = 20,656; with only three indicators the model cannot be tested; χ^2 = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; CFit = close fit (probability that RMSEA < .05); CFI = comparative fit index. SRMR = standardized root mean square residual.

Research Question 2: What evidence exists to support a higher-order factor structure of teacher effectiveness, at both the student and classroom level? To further the exploration of factor structure at the classroom level, I also fit several higher-order confirmatory factor models, which are also referred to as hierarchical factor models. Higher-order factors are hypothesized to influence or explain variation in the lower order factors in a specific manner (Brown, 2015). Unlike the confirmatory models analyzed in the previous section, here I specifically model the interrelation of the seven factors. This aspect of the analysis reflects (a) configurations implied by the Tripod authors (Models 5, 6, and 7) and (b) a bifactor structure introduced by Wallace and coauthors (Models 8, 9 10, and 11) as an alternative specification to account for variation among the seven factors (Wallace et al., 2016). The four bifactor models posit a general factor representing teacher effectiveness to explain a significant portion of the variance among the observed indicators. In addition to this general factor, additional factors are also included to account for unique variation beyond what is captured by the general factor.

As in the previous section, I ignore the multilevel structure (i.e., the nesting of students within classroom, of the data and estimate seven models employing a cluster correction to account for nonindependence in the response data through adjusted standard errors. I also evaluated the same suite of fit indices to evaluate the models with the goal of determining the best model or models to bring forward to the multilevel analysis. I begin the analysis with the Tripod-inspired higher-order models. These models hypothesize that the variance in the seven first-order factors is accounted for by a more general, higher-order factor. Model 6 represents a single second-order factor to accounts

for the variation among the seven first-order factors. In contrast, Model 7 specifies three second-order factors to explain the variation in specific first-order factors. The final Tripod motivated model, Model 8, adds a single third-order factor to Model 7 to account for the variation among the three second-order factors. (See Figure 2.)

All three of the Tripod-based models (Models 6–8) exhibited roughly the same level of fit with RMSEA = .061, CI for $p < .05 = 0$, SRMR = .046, and a range for CFI of .864–.867. (See Table 7.) These results are similar to the Tripod-based first-order models evaluated in the previous section in that they also exhibited mixed results. While values for RMSEA and SRMR fall within acceptable ranges, the values for CFI are well both below the common, minimum cutoff of .90 and strongly suggesting model miss-specification (Brown, 2014; Kline, 2015). An examination of possible areas of misfit within the modification indices also indicated a high number of cross-loadings among indicators and factors. This result is not altogether unexpected, given the potential misspecification of the first-order factors models. An examination of the residuals, however, yielded no unusual results except that Model 8 resulted in a negative residual variance on the second-order factor of Curricular Support. Negative residual variances are also an indication of model miss-specifications (Brown, 2014).

Figure 12 depicts the final set of models tested here that describe a bifactor structure. In these types of models, a general factor accounts for a significant degree covariation among the observed indicators together with one or more additional factors specified to explain variance above and beyond the general factor. Model 9 represents a general factor for teacher effectiveness and seven subdomain factors (i.e., the 7Cs).

Model 10 posits a general factor and three subdomain factors of curriculum support, personal support, and academic support. Model 11 represents a bifactor specification developed Wallace et al. (2016), who demonstrated moderate levels of fit with data drawn from the MET data in Grades 6–8. The Wallace model hypothesizes a general factor with two additional factors; one representing the classroom control factor and a second methods factor representing the negatively worded items within the Tripod survey. Of note, while Models 9 and 10 are somewhat parallel to the second-order Tripod models above, the Wallace specification represents an alternative formulation of teacher effectiveness that the authors found to be theoretically and empirically sound (Wallace, 2016).

I first fit two bifactor alternatives of the Tripod utilizing two models. Model 9 represented a general factor and seven subdomain factors. Model 10 represented a general factor and three subdomain factors. In both models, all 36 items load onto the general factor and subsets of the items also load onto the domain-specific factors (see Figure 12). While Model 9 failed to converge, Model 10 exhibited similar levels of fit to the second-order models (Models 5–7) across the suite of fit indices (RMSEA = .061, CI for $p < .05$ = 0.00, CFI = .864; SRMR = .046).

I next fit the Wallace specification (Model 11) with acceptable to good levels of model fit (RMSEA = .049, CI for $p < .05$ = .962, CFI = .912, SRMR = .032). The final model in the bifactor series is a modified version of the Wallace specification that incorporated two well-fitting factors from the Tripod structure, Clarify and Captivate as a substitute for the Control factor. I include these two factors because they emerged in my

initial analysis as the only individual factors (out of the original seven) with acceptable levels of model fit (see Table 8). I also kept Wallace's method factor consisting of the two negatively worded items. The results for Model 12 were strong, demonstrating slightly improved fit over Model 11 (RMSEA = .044, CI for $p < .05$ = 1.00, CFI = .931; SRMR = .029). See Figure 12 for a diagram of Models 9–12.

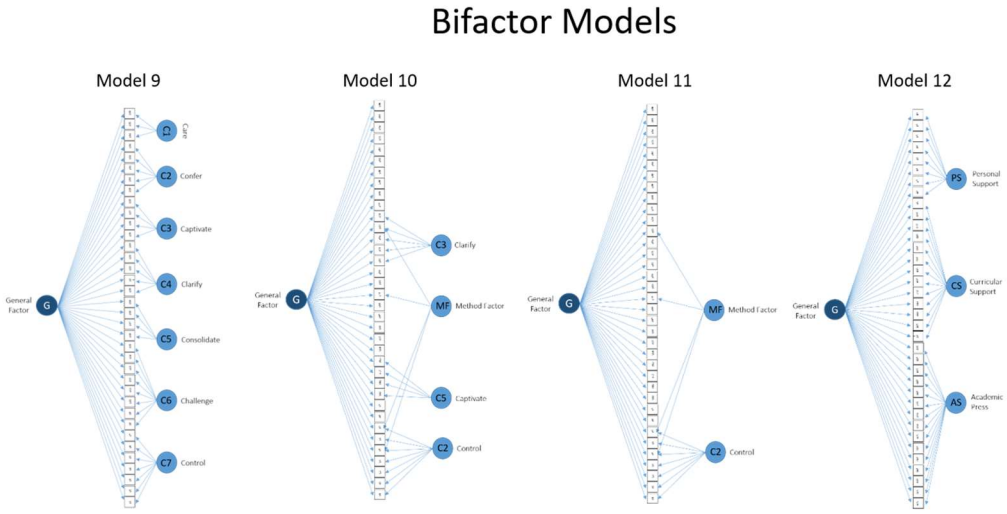

Figure 12. Four confirmatory bifactor models.

To summarize, the purpose of RQ2 was to finalize Step 1 in the multilevel confirmatory factor analysis strategy by determining if a higher-order factor structure might further explain variation in the observed indicators beyond the first-order models estimated in the previous section (RQ1). In the higher-order category of models, I tested seven models including three models representing second and third order factor structures for the Tripod survey (Models 6, 7, and 8), as well as three bifactor models (Models 9–12). Only two of the bifactor models achieved good levels of model fit, the Wallace (Model 11) and hybrid Wallace (Model 12).

Table 8

Year 1 Test of Model Fit: Higher Order Factor Structure

Models		χ^2	df	RMSEA (90% CI)	CFit	SRMR	CFI
Hypothesized Factor Structure							
Model 6	1-2nd Order Factor	44,228.045	587	.061 (.06-.061)	0.00	.046	.867
Model 7	3-2nd Order Factors	45,174.809	584	.061(.061-.062)	0.00	.046	.864
Model 8	1-3rd 3-2nd Order	45,173.445	584	.061 (.061-.062)	0.00	.046	.864
Bifactor Models							
Model 9	7C Bifactor	Did not converge					
Model 10	3C Bifactor	45,174.809	584	.061(.061-.062)	0.00	.046	.864
Model 11	Wallace Bifactor	29,422.351	582	.049(.049-.050)	.962	.032	.912
Model 12	Wallace2 Bifactor	23,172.131	573	.044(.044-.045)	1.00	.029	.931

Note. n = 20,656; Correction to account for the non-independence of the survey items. χ^2 = chi-square; *df* = degrees of freedom; RMSEA = root mean square error of approximation; CFit = close fit (probability that RMSEA < .05); CFI = comparative fit index. SRMR = standardized root mean square residual.

These two models, together with the three models identified in the previous section, provided the strongest candidate models to conclude Step 1 and bring forward into the next step in the multilevel strategy. The noted exception is the ESEM specification, Model 4, which demonstrates excellent fit; however, because of the lack of computational routines for the multilevel version of the ESEM specification, I could not estimate a multilevel version of this model. Also of note is the seven-factor Tripod model, Model 3, which performed worse that the other three models. However, given the theoretical importance of the Tripod seven-factor model, its widespread use in school districts throughout the United States, and its ambiguous model fit, added it to the slate of candidate models to test in the multilevel environment. In the following section I first evaluate the degree of L2 variation and then estimate the multilevel versions of Models 3, 6, 11, and 12.

Research Question 3: Does teacher effectiveness, as measured by the Tripod survey, represent a multilevel construct? As outlined previously, a single-level analysis is the first step in a process to evaluate multilevel models. In this section, I incorporate Step 2 and Step 3 of the multilevel confirmatory factor analysis strategy by first evaluating (a) the classroom-level variation and second, (b) estimating the level-specific versions the four models from Step 1. The evaluation can determine if any of the single-level models exhibit the same factor structure at the classroom level (i.e., configural isomorphism.

Classroom-level variation. To evaluate classroom level variance, I used intraclass correlation coefficients, (ICC1 and ICC2) and the DE. Note that I use *between* to describe the group or classroom level of analysis and *within* to describe the individual or student level of analysis. ICC1 is the ratio of the between level variance and the total variance for an observed variable with values ranging from 0.00 to 1.0 (Byrne & van de Vijver, 2014). As an example, Item 1 of the Tripod Survey ("My teacher in this class makes me feel that h/she really cares about me") has an ICC1 of 0.203. (See Table 9.) For this item, the amount of variance explained by group membership is roughly 20% and indicates the variance accounted for by student membership in a particular classroom. The ICC1 values for all items ranged from 12–31%. Item 15 ("When h/she is teaching us my teacher thinks we understand even when we don't") exhibited the least amount of group influence, ICC1 = .12 and Item 6 ("Students in the class treat the teacher with respect") exhibited the greatest degree of group influence, ICC1 = .32. Within the multilevel factor analysis literature, ICC1 values greater than .05 provide evidence of group-level variances in observed indicators, and the hierarchical structure of the data should not be ignored (Brown, 2014; Kline, 2015; Muthén, 1991)

The second type of interclass correlation, ICC2, is derived from ICC1, and the mean class size and is interpreted as a measure of reliability for the group mean of each item. ICC2 ranged from 67.2% to 88.5% with Item 15 exhibiting the lowest level reliability and Item 6 exhibiting the highest level of reliability. ICC2 values can be evaluated from poor to excellent depending on their magnitude. Values less than .7 are

considered poor, .7 to .8 are considered acceptable, .8 to .9 are considered good, and greater than .9 are excellent (Bliese, Maltarich, Hendricks, Hofmann, & Adler, 2019).

I also considered the DE as a measure of group level variance. The DE is a function of ICC1 and cluster (group) size and indicates the degree of non-independence at the group level. The DE ranged from 2.71 to 5.99, with Item 15 exhibiting the smallest DE and Item 6 with the largest. Design effects greater than 2.0 for individual items are considered appropriate for multilevel analysis. Not surprisingly the DE results are consistent with the ICC1 and ICC2 values mentioned above and provide a measure of the differences in student perceptions of individual items (Bliese et al., 2019).

To summarize, the ICCs and DE measures for the Tripod items provided strong evidence of classroom level variation. ICC1s ranged from 12–31%, suggesting a high degree of group variation. ICC2 ranged from .672 to .885 suggesting acceptable to good degrees of reliability for the group means. Finally, DEs were well above 2.0, indicating a high degree of dependence within groups. Taken together, these values clearly reflect more than adequate variance at the between level (the classroom clusters), providing ample evidence of a multilevel construct.

Table 9

Year 1 Tripod Items Intraclass Correlations and Design Effects

Item	Within	Between	ICC (1)	ICC(2)	Design Effect
1	1	0.255	0.203	0.813	4.272
2	1	0.251	0.201	0.811	4.231
3	1	0.198	0.165	0.772	3.661
4	1	0.247	0.198	0.809	4.190
5	1	0.342	0.255	0.854	5.104
6	1	0.449	0.310	0.885	5.990
7	1	0.316	0.240	0.844	4.867
8*	1	0.238	0.192	0.803	4.096
9*	1	0.369	0.270	0.863	5.340
10*	1	0.372	0.271	0.864	5.366
11	1	0.208	0.172	0.781	3.773
12	1	0.239	0.193	0.803	4.106
13	1	0.233	0.189	0.799	4.043
14	1	0.164	0.141	0.737	3.269
15*	1	0.12	0.107	0.672	2.725
16	1	0.149	0.130	0.718	3.088
17	1	0.182	0.154	0.757	3.479
18	1	0.144	0.126	0.711	3.027
19	1	0.163	0.140	0.736	3.257
20	1	0.192	0.161	0.767	3.594
21	1	0.191	0.160	0.766	3.582
22	1	0.169	0.145	0.743	3.328
23	1	0.145	0.127	0.713	3.039
24	1	0.373	0.272	0.864	5.375
25	1	0.341	0.254	0.854	5.095
26	1	0.315	0.240	0.843	4.857
27*	1	0.171	0.146	0.745	3.352
28	1	0.191	0.160	0.766	3.582
29	1	0.24	0.194	0.804	4.117
30	1	0.22	0.180	0.790	3.904
31	1	0.232	0.188	0.799	4.032
32	1	0.178	0.151	0.753	3.433
33	1	0.204	0.169	0.777	3.728
34	1	0.196	0.164	0.770	3.639
35	1	0.179	0.152	0.754	3.445
36	1	0.234	0.190	0.800	4.054

Note. * indicates a reverse-coded item.

Level-specific model fit. Given the degree of between-level variation discussed above, I next evaluated the multilevel versions of the four models identified from RQ1 and RQ2. In a typical analysis of multilevel models, the entire model hypothesized model is simultaneous estimated by using the sample variance-covariance matrix at each level. The results of this analysis are then evaluated to determine model fit. However, this approach routinely masks model misfit at the between level (Rhu et al., 2009), because the much larger sample size at the within level dominates the fit statistic calculation (Stapleton et.al., 2016). In my analysis, the sample size at the individual level is roughly 17 times larger than the sample size of the classroom level (N_w = 20,656, N_b = 1,188). This dominance by the within-level fit makes it difficult to differentiate model fit between levels (Janis et al., 2016; Ryu & West, 2009). To overcome this shortfall, I employed a level-specific approach that uses partially saturated models to determine fit at both levels. This alternative approach, developed by Ryu and West (2009), provides a more precise method for detecting model misfit at each level and facilitates proper identification of good-fitting multilevel measurement models (e.g., configural isomorphism). As noted previously, determining the degree of model fit is of particular importance when a multilevel measurement model is used in a subsequence analysis (Ryu, 2014; Ryu & West, 2009; Stapleton et al., 2016).

I began this section by estimating a series of partially saturated submodels to separately determine the model fit at the individual (within) and group (between) levels. To begin, I first specified three types of submodels: (a) saturated, (b) independent, and (c) hypothesized model (both within and between). A saturated model, by definition,

provides perfect fit to the data by utilizes all possible variances and covariances of the observed data. This produces the exact data-implied variances-covariance matrix with zero degrees of freedom. In contrast, the independent model, by definition, provides the poorest fit to the data by specifies zero correlations between the items. This model then provides a baseline for the worst-fitting model to the data. The saturated and independent models are necessary inputs for the calculation of model fit indices. The hypothesized models are also estimated and yield fit results that fall somewhere between the perfect and poorest model fit (saturated and independent models, respectively). With these submodels in mind, the partially saturated strategy consists of fitting one level as saturated while the other level is specified as the hypothesized factor structure. This provides evaluation of model fit at the hypothesized level, thus constraining any potential misfit to the hypothesized level only because saturated-level contains perfect fit. The model chi-squared statistics generated by the submodel then reflects the non-saturated level and can be used to calculate the level-specific model fit statistics of RMSEA and CFI (Ryu & West, 2009).

To organize this analysis, I show the four specifications needed to determine fit at the between and within levels of each model in Table 10. For each of the models identified as appropriate at the individual level, I then fit the four submodels (a–d). Each submodel has a between-level specification and within-level specification.

Table 10

Partially Saturated Model Specification

Models	Between	Within
a	Saturated	Hypothesized
b	Saturated	Independent
c	Hypothesized	Saturated
d	Independent	Saturated

Consistent with previous sections, I utilized RMSEA and CFI as model fit indices for the within and between levels. Based on the above, I reported model chi-square, degrees of freedoms, and sample size for the 16 submodel (four models times four submodel), in Appendix B. These values were used to calculate model fit statistics for both levels of each model. The formulas, per Ryu and West (2009) for both indices, follow:

Between-Level Fit Statistics:

$$RMSEA_b = SqRt\ (\chi^2_{H0bSATw} - df_{H0bSATw}) \quad (1)$$

$$CFI_b = 1 - (\chi^2_{H0bSATw} - df_{H0bSATw}) / (\chi^2_{INDbSATw} - df_{INDbSATw}) \quad (2)$$

Within-Level Fit Statistics:

$$RMSEA_w = SqRt\ (\chi^2_{SATbH0w} - df_{SATbH0w}) \quad (3)$$

$$CFIw = 1 - (\chi^2_{SATbH0w} - df_{SATbH0w}) / (\chi^2_{SATbINDw} - df_{SATbINDw}) \quad (4)$$

Where $\chi^2_{H0bSATw}$ is a model with the hypothesized model specified at the between level and the saturated model is specified at the within level and *df* are the degrees of freedom.

Using the level-specific indices in Equations 1–4 above, I then estimated all the submodel for Models 3, 5, 11, and 12 followed by calculations of level-specific fit indices. Table 11 provides a summary of three results for each model and includes (a)

overall hypothesized multilevel model fit for the combined between and within levels (H_0-H_0), (b) the hypothesized model between, saturated within (H_0-SAT), and (c) saturated between, hypothesized within (SAT-H_0).

Beginning with Model 3, the Tripod seven-factor model, I first estimated the hypothesized seven factor model with identical structures at both the individual and group levels. This specification is essentially assuming configural equivalence across both levels. Described as the "Overall" model in Table 9, the estimates of fit statistics yielded mixed results (χ^2 = 70,949.07 (1146), RMSEA = .055, CFI = .758, and SRMRb = .049, SRMRw = .054). Although RMSEA and SRMR at both levels appears to indicate good fit, a CFI of .758 strongly suggests miss-specification with the configural assumption (equal facture structures at both levels). Turning to the level-specific fit indices provides insight into the specific location of misspecification. While the within-level indices are adequate ($RMSEA_w$ = .087, CFI_w = .972) the between-level model is clearly mis-specified ($RMSEA_b$ = .138, CFI_b = .919) as RMSEA values > .1 are considered uninterpretable.

The results for the two-factor model, Model 5, demonstrated a very different outcome. The overall model fit (H_0H_0) exhibited very good (χ^2 = 1667.934 (51), RMSEA = .04, CFI = .985, and SRMRb = .008, SRMRw = .019), as did the level-specific fit indices, ($RMSEA_w$ = .053, CFI_w = .997, $RMSEA_b$ = .035, CFI_b = .995). This outcome strongly suggests an equal factor structure across levels and thus configural isomorphism.

I next examined the two bifactor models, Model 11 and Model 12. Note that the overall fit for both models also exhibited ambiguous and mostly similar levels of fit to

Model 3. The Wallace Bifactor, Model 11, resulted in χ^2 = 40,989.31(1164), RMSEA = .041, CFI = .862, and SRMRb = .055, SRMRw = .033, and for the hybrid-Wallace, Model 12 resulted in χ^2 = 31,586.01.31(1146), RMSEA = .036, CFI = .895, and SRMRb = .060, SRMRw = .030. An examination of the level-specific fit indices also indicted a similar pattern to Model 3, with misfit located at the between level ($RMSEA_b$ = .141, CFI_b = .914) for Model 11 and very similar values for Model 12 ($RMSEA_b$ = .135, CFI_b = .917). Both Model 11 and Model 12 performed well at the within or individual level ($RMSEA_w$ = .063, CFI_w = .985, $RMSEA_w$ = .055, CFI_w = .989). These results provide compelling evidence that the hypothesized Tripod factor structure does not hold across levels due to poor model fit at the between-level and thus does not meet the criteria for configural isomorphism. Further, the bifactor versions of the Tripod also do not meet the standard for configural isomorphism. These results call into question any conclusions drawn from differences at the classroom level.

To summarize RQ3, I tested a series of partially saturated models to evaluate level-specific fit for four multilevel models. Only Model 5 demonstrated equal structure across levels as evidenced by the level-specific fit indices. Models 3, 11, and 12 all demonstrated a high degree of misspecification at the between-level, indicating that the factor structures of the Tripod and the two bifactor models are not configurally equivalent to their within-level counterparts. This lack of equal factor structures across levels has consequences for interpreting group-level variance based on individual, student-level rating (Jebb, Tay, Ng, & Woo, 2019; Tay, Woo, & Vermunt, 2014). The relatively good fit of these models at the individual level captures student differences but disqualifies

these three models (Models 3, 11, and 12) as measurement tools that differentiate a group level attribute, which is a prerequisite for comparing factor variances at the group level (Stapleton et al., 2016)

The two-factor model (Model 5), however, exhibited excellent fit to the data, providing strong evidence of equal factor structure across levels (i.e., configural isomorphism). This result makes Model 5 a candidate for metric isomorphism across both levels (i.e., equal factor loadings). Recall, both configural and metric isomorphism are required for valid comparisons of group level differences (Mehta & Neale, 2005; Tay et al., 2014). In the next section I test for metric isomorphism of the two-factor model (Model 5) by specifying equal factor loadings across Levels 1 and 2.

Research Question 4: To what extent does the construct of teacher effectiveness, as measured by the Tripod survey, hold across levels (individual and aggregated)? In the previous section I tested four multilevel models for configural isomorphism (Step 3 in the multilevel confirmatory factor analysis process), using partially saturated submodels to identify model misfit at the individual and group levels (within and between). Only one model achieved strong configural equivalence across levels, the two-factor reduced form Tripod with the constructs of Clarify and Captivate (Model 5). In this section, I proceed to the final step in establishing multilevel isomorphism, testing the metric invariance (e.g., equivalent factor loadings across levels) of Model 5. Testing metric invariance across levels provides evidence that the scale being used to measure teacher effectiveness from the individual level can be aggregated to the classroom level for comparisons across groups (Mehta & Neale, 2005). If, on the other

hand, metric invariance does not hold, then the aggregated student ratings are not valid comparisons of teacher effectiveness at the classroom level.

Metric invariance. To evaluate the property of metric invariance for Model 5, I fit three additional specification, each designed to systematically test for metric invariance. Specification 5a represents the equal factor structure across levels from the previous section (i.e., configural isomorphism). In Specification 5b, I constrain the residual variances at the between level to zero. This constraint forces all variation in the items to be accounted for by the two common factors. Any decrement in model fit between 5a and 5b indicates a degree of variation not accounted for by the common factors. In Specification 5c, I add the additional constraint of equal factor loadings. In other words, in 5b all factor loadings are freely estimated at both levels and are expected to result in similar magnitudes across levels but with slightly different values (e.g., λ_{14w}= 749.and (λ_{14b} = .766). In Specification 5c, these factor loadings fixed to equality ($\lambda_{14w} = \lambda_{14b}$). Equal factor loadings across levels makes explicit the concept of measurement equivalence at both levels (i.e., metric isomorphism). And to complete this series, I free the residual variances in 5c to evaluate any changes in model fit and name this Specification 5d, see Figure 13.

Table 11

Level Specific Test of Model Fit

Models	Description	χ^2	*df*	*p*	RMSEA	CFI	SRMRw	SRMRb
Model 3	7-factor							
Overall		25702.560	1146	0.00	.051	.771	.748	.048
Classroom-level		5073.289	573	0.00	.137	.928	-	-
Student-level		35848.444	573	0.00	.087	.919	-	-
Model 5 RF*	2-factor							
Overall		1653.150	50	0.00	.040	.985	.019	.008
Classroom-level		98.936	25	0.00	.0499	.986	-	-
Student-level		1226.56	25	0.00	.053	.973	-	-
Model 11	Wallace bifactor							
Overall		40989.309	1164	0.00	.041	.862	.033	.055
Classroom-level		14244.829	582	0.00	.140	.914		
Student-level		44467.201	582	0.00	.063	.985		
Model 12	Wallace hybrid							
Overall		31,586.006	1146	0.00	.036	.895	.03	.06
Classroom-level		13767.660	573	0.00	.135	.917	-	-
Student-level		3344.561	573	0.00	.055	.989	-	-

Note. N_w = 20,656, N_b = 1,188; *RF = residual factor to account for cross loadings of the two reverse coded items (R_C15 with R_C27). χ^2 = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; CFit = close fit (probability that RMSEA < .05); CFI = comparative fit index; $SRMR_w$ = within level standardized root mean square residual; $SRMR_b$ = between level standardized root mean square residual.

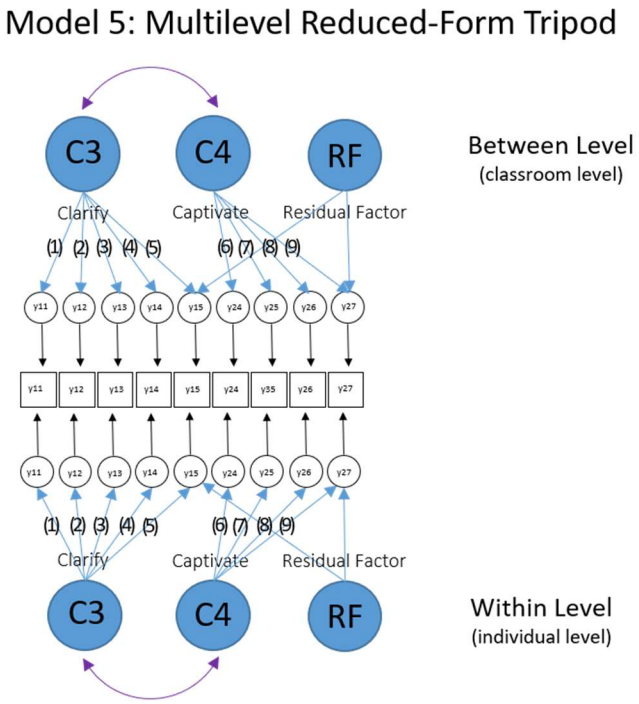

Figure 13. Model 5d with factor loadings. Loadings with the same number in parentheses are equal across levels.

In Table 12, I provide model fit information for the four specification of Model 5 (a–d). As described above, these specifications are designed to aid in evaluating potential areas of model misfit and provide a systematic process to determine metric invariance (add citation). Model 5a, as previously reported, provided strong fit to the overall data, χ^2 = 1653.15 (50), RMSEA = .04, CFI = .985, and SRMRb = .008, SRMRw = .019. In Model 5b, with residual variances constrained to zero at the between-level, also provided good fit, although with some decrement (χ^2 = 1816.484 (61), RMSEA = .06, CFI = .957, and SRMRw = .033, SRMRb = .039). I next fit Model 5c, with factor loadings constrained to be equal across levels. The results indicated continuing good fit to the data with χ^2 = 1042.679 (66), RMSEA = .028, CFI = .990, SRMRw = .019, and SRMRb = .038. I also fit a final model in this sequence in order to evaluate the assumption of zero

residual variances at the between level. I accomplish this by removing the constraint of zero residual variances. Model 5d demonstrated very similar results with χ^2 = 1138.533 (57), RMSEA = .031, CFI = .990, and SRMRw = .0019, SRMRb = .035. These results provide strong evidence of metric invariance across levels for Model 5 and supports the argument for valid comparisons of differences between classrooms based on the two-factor reduced form Tripod model.

Table 13 provides parameter estimates and standard errors of Model 5c (i.e., metric equivalence). I calculate the proportion of variance in the two common factors of Clarify and Captivate at the classroom level as

$$\sigma_B^2 \,/\, (\sigma_B^2 + \sigma_W^2)$$

The proportion of variance in the factor Clarify, at the classroom level is .371 / (1.00 + .371) = .271 and for Captivate is .486 / (1.00 + .486) = .327. This indicates that 27.1% of the variance in the teacher quality, as measured by the factor Clarify, exists between classrooms. By definition, 72.9% of the variation in the items measuring Clarify exists at the student level. For the second factor, Captivate, 32.7% of the variance exists at the classrooms level with the remaining 67.3% at the student level.

Table 12

Year 1 Model 5 Test for Equality of Measurement Structure

Models	Description	χ^2	*df*	*p*	RMSEA	CFI	SRMRw	SRMRb
Model 5a	Loadings + residuals estimated	1653.150	50	0.0	.040	.985	.019	.008
Model 5b	Residuals @0	1816.484	61	0.0	.060	.957	.033	.039
Model 5c	Equal loadings, residuals @0	1110.511	66	0.0	.028	.990	.019	.038
Model 5d	Equal loadings, residuals estimated	1138.533	57	0.0	.031	.990	.019	.035

Note n_w = 20,656, n_b = 1,188; χ^2 = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; CFit = close fit (probability that RMSEA < .05); CFI = comparative fit index; $SRMR_w$ = within level standardized root mean square residual; $SRMR_b$ = between level standardized root mean square residual.

Table 13

Year 1 Parameter Estimates for the Two-Factor Reduced Form Model

	Within				Between			
	Est.	*SE*	*p*	Std.	Est.	*SE*	*p*	Std.
				Factor 1				
C11	1.203	.014	0.0	.769	1.201	.014	0.0	1.024
C12	1.331	.016	0.0	.799	1.331	.016	0.0	.992
C13	1.227	.014	0.0	.775	1.228	.014	0.0	.974
C14	.903	.011	0.0	.670	.903	.011	0.0	1.004
R_C15	.565	.009	0.0	.449	.565	.009	0.0	.785
				Factor 2				
C24	1.769	.020	0.0	.871	1.769	.020	0.0	.998
C25	1.559	.016	0.0	.841	1.559	.016	0.0	1.00
C26	1.125	.016	0.0	.747	1.125	.016	0.0	.923
R_C27	.800	.010	0.0	.579	.800	.010	0.0	.970
				Residual factor				
R_C15	1.00			.410	1.00			.487
R_C27	1.00			.374	1.00			.326
				Factor variances				
Factor 1	1.00	.00	999	1.00	.368	.018	0.0	1.00
Factor 2	1.00	.00	999	1.00	.480	.024	0.0	1.00
RF	.267	.011	0.0	1.00	.007	.009	0.0	1.00
				Factor covariance				
$Cov_{(F1, F2)}$	.880	.003	0.0	.880	.400	.020	.00	.937
$Cov_{(F1, RF)}$	.00	.00	999		.00			.00
$Cov_{(F2, RF)}$	.00	.00	999		.00			.00

Note n_w = 20,656; Est. = factor loading estimate; *SE* = standard errors; *p* = probability level of .05; Std. = standardized estimate; *RF = residual factor to account for cross loadings of the two reverse coded items (R_C15 with R_C27).

To summarize, this section, I further developed the case for isomorphism of the two-factor reduced form Tripod model (Model 5a) by constraining the factor loading loadings to be equal across levels (Model 5c). This model performed well, thus providing strong evidence for metric invariance of the two-factor model (χ^2 = 1042.679 (66), RMSEA = .028, CFI = .990, and SRMRw = .019, and SRMRb = .038). I also tested the

assumption of zero residual variances by fitting Model 5d with freely estimated residual variances which resulted in very similar levels of model fit. Based on these results, I concluded that strong metric invariance holds across classrooms for these items. Finally, with metric invariance established, I calculated the variances for the two model factors of Clarify and Captivate. This calculation demonstrated that considerable differences in the teacher practice exists between classrooms on these two dimensions of teacher effectiveness (27.1% and 32.7%, respectively).

Research Question 5: Does the Tripod survey exhibit a consistent factor structure across measurement periods? In this final section, I evaluate Year 2 of the MET data to confirm the findings from Year 1. In Year 2, 8,111 students were clustered into 419 classrooms. The teachers in Year 2 represented a subset of the original Year 1 teachers. As can be surmised from the different sample sizes, about two thirds of the original Grade 9 teachers dropped out of the study in the second year. (See Table 3.) The descriptive statistics for the 36 items of the Tripod Survey in Year 2 can be found in Table 4. Also of note, in the Year 2 sample, is the randomization of students into sections, contrasted with the Year 1 sample, wherein student assignments to individual class sections were not controlled.

Year 2: Single-level models. I began the analysis of the Year 2 data by confirming results from Year 1, including (a) estimating the individual factors of the Tripods survey (Care, Control, Clarify, Challenge, Captivate, Confer, and Consolidate); (b) estimating all twelve models from RQ1 and RQ2. The results of the individual factor analysis revealed general similarities to the Year 1 sample. The individual factors of

Control and Consolidate also demonstrated the poorest fit in the Year 2 sample. Challenge and Confer also demonstrated marginal results, although Confer did exhibit some improvement in fit over Year 1. The strongest performing factors in Year 1, Clarify and Captivate, did continue to demonstrate the best fit in Year 2, even though both exhibited some reduction in RMSEA over Year 1. These results can be found in Appendix C.

In addition to the individual factors, I also revisited the full set of Year 1 models (Models 1–12) in relation to the Year 2 data to confirm my previous results. These models were divided into two groups and represented a systematic process to determine the best fitting multilevel model in the Year 1. The first group of models represented the single-order factor structures (Models 1–5). Of particular interest was Model 3, the hypothesized seven factors of the Tripod Survey. This model provided highly ambiguous results in the Year 1 data and exhibited similar results in Year 2, with good fit on the RMSEA index but poor fit via the CFI index. The best fitting models from this group (i.e.,,, single-level, first order) in Year 2 were also the top models in Year 1 (the seven-factor ESEM and the two-factor reduced form of the full Tripod, Models 4 and 5, respectively). These two models demonstrated nearly identical results across both sample periods. (See Appendix D for full results.)

I next highlight the second set of models representing the higher-order factor structures described in detail in section RQ2 (Models 6–12). The results from the Year 2 sample again provided similar results to Year 1: The Tripod based high-order models exhibited highly ambiguous fit or did not converge. The two best fitting models in Year 2

were also the bifactor models based on a specification proposed by Wallace et al. in 2018 and corresponded to the same models with the best fit in Year 1. To summarize, the Year 2 analysis of the single-level models was identical to Year 1, with Models 5, 11, and 12 exhibiting the most consistent and acceptable levels of fit. (See Appendix C.)

The primary motivation for exploring single-level models in multilevel settings is to establish a good fitting level-one model to bring forward into a fully multilevel analysis (Heck & Thomas, 2015; Muthén, 1991). As established with RQ 2 and RQ3, Models 5, 11 and 12 represented the most appropriate models for multilevel analysis because of their nonambiguous model fit. Model 4, the ESEM model, also provided excellent fit to the Year 2 data; however, as stated earlier, no methodology exists to extend this configuration to the multilevel environment. Thus, as in the Year 1 analysis, I provide no further information regarding this model. Having established a likely set of candidate models for multilevel modeling in Year 2, I also briefly discuss the degree of variation found at the group level of data, which is a prerequisite for conducting multilevel analysis. Following the same process in the section on RQ3, I highlight the ICCs and the DE before proceeding to review the multilevel results.

Year 2 classroom variability. The degree of variability at the classroom level, as measured by the average intraclass correlation (ICC1), was 17.5% in Year 2, with a low of 10.1% on Item 8 (“In this class, my teacher accepts nothing less than our full effort”) to a high of 24.8% on Item 6 (“Students in this class treat the teacher with respect”). This compares to an average ICC1 of 18.6% in Year 1. Recall that ICC2 measure is based on the magnitude of ICC1 and the average classroom size, and it is considered a measure of

reliability for group means. The average ICC2 in Year 2 was 77.9% as compared to 78.8% in Year 2. The final measure of group level variability is the DE, which is also a function of ICC1 and the average cluster size and quantifies the degree to which the sample is cluster dependent. Values great than 2.0 are generally considered to exhibit variability at the cluster level. The Year 2 DE was equal to 3.81 as compared to 4.0 in Year 1. Taken together, all three measures demonstrate a high degree of cluster level variability, providing evidence for a multilevel analysis of the Year 2 data. Measures are reported in Table 13.

Year 2 level-specific test of model fit. Given the degree of cluster level variance, I next to proceed to the multilevel analysis and evaluate the three best fitting single-level models from Year 2 (Models 5, 11, and 12). As in the Year 1 analysis, I also fit Model 3, even though the fit was highly ambiguous because of its prevalence in U.S. school districts (Geiger & Amrein-Beardsley, 2019). Table 14 provides overall model fit, as well as level specific indices for Models 3, 5, 11, and 12. Examining the overall model variance-covariance matrix provided nearly identical results between Years 1 and 2. As in Year 1, both Models 3 and 11 exhibit highly mixed results when examining goodness of fit measures (Model 3: RMSEA = .051, CFI = .771; Model 11: RMSEA = .037, CFI = .877), while Models 5 and 12 exhibited more consistent results (Model 5: RMSEA .062, CFI = .961; Model 12: RMSEA = .032, CFI = .909).

Table 14

Year 2 Tripod Items Intraclass Correlations and Design Effects

Item	Within	between	ICC (1)	ICC(2)	Design Effect
1	1	0.303	0.233	0.838	4.745
2	1	0.267	0.211	0.820	4.393
3	1	0.223	0.182	0.792	3.936
4	1	0.179	0.152	0.754	3.445
5	1	0.219	0.180	0.789	3.893
6	1	0.276	0.216	0.825	4.483
7	1	0.175	0.149	0.750	3.398
8	1	0.112	0.101	0.657	2.622
9	1	0.202	0.168	0.776	3.706
10	1	0.178	0.151	0.753	3.433
11	1	0.225	0.184	0.794	3.958
12	1	0.246	0.197	0.808	4.179
13	1	0.241	0.194	0.805	4.127
14	1	0.192	0.161	0.767	3.594
15	1	0.181	0.153	0.756	3.468
16	1	0.156	0.135	0.727	3.173
17	1	0.237	0.192	0.802	4.085
18	1	0.172	0.147	0.746	3.363
19	1	0.167	0.143	0.741	3.304
20	1	0.181	0.153	0.756	3.468
21	1	0.212	0.175	0.784	3.817
22	1	0.168	0.144	0.742	3.316
23	1	0.15	0.130	0.720	3.100
24	1	0.329	0.248	0.849	4.986
25	1	0.305	0.234	0.839	4.764
26	1	0.271	0.213	0.823	4.433
27	1	0.204	0.169	0.777	3.728
28	1	0.236	0.191	0.801	4.075
29	1	0.222	0.182	0.792	3.925
30	1	0.209	0.173	0.781	3.784
31	1	0.229	0.186	0.797	4.000
32	1	0.2	0.167	0.774	3.684
33	1	0.216	0.178	0.787	3.860
34	1	0.216	0.178	0.787	3.860
35	1	0.168	0.144	0.742	3.316
36	1	0.229	0.186	0.797	4.000

As outlined previously, an examination of the level-specific model fit provides a clearer picture of multilevel models. A major drawback of fitting the full variance-covariance matrix is that Level 1 data tend to dominate the fit indices because of the much larger sample size at level one, which can lead to misinterpretation of model output. Using Rhu and West's (2009) partially saturated methodology, I fit submodel representing combinations of the (a) hypothesized, (b) saturated, and (c) independence models at each level (see Table 10 for specifications). For the classroom-level only models (hypothesized factor structure at L2 and saturated factor structure at L1), the corresponding RMSEA and CFI indices provided consistent evidence of misfit for Models 3, 11, and 12. (See Table 14.) In addition, these results also replicate the conclusion from Year 1: These models do not exhibit configural invariance and the misfit is located at the classroom level. Further, these results disqualify these models from further analysis and are not valid measures of the classroom level constructs they are purported to measure.

While the findings from Models 3, 11, and 12 were highly consistent with Year 1, what was different in Year 2 was the ambiguous fit for Model 5. While the results for the Year 2 overall model were strong, $RMSEA_{O2}$ = .044 and CFI_{O2} = .981 at the student level, there was considerable degradation in model fit on $RMSEA_{W2}$ = .089 as compared to Year 1 $RMSEA_{W1}$ = .053 and to a lesser degree on CFI_{W2} = .970 as compared to Year 1 CFI_{W1} = .992. This degradation was also present at the classroom level, with $RMSEA_{B2}$ = .095 as compared to Year 1 $RMSEA_{B1}$ = .05 and again to a lesser degree for CFI_{B2} = .981 as compared to Year 1 CFI_{B1} = .986. Taken together, these results suggest

differences in the configuration of pattern of item loadings across levels for Model 5 as well. (See Table 14 for full result of this analysis.)

To further explore the decrement in fit for Model 5, I take advantage of a variable in the dataset distinguishing the three classroom subjects taught in the Grade 9 (subject = Algebra I, Biology, English Language Arts). The sample sizes for the individual responses in Year 2 were $N_{(Alg1)}$ = 2,351, $N_{(Bio)}$ = 2671, and $N_{(Ela)}$ = 3065, for a total sample size = 8,087. The number of sections by subject in Year 2 was $N_{(Alg1)}$ = 132, $N_{(Bio)}$ = 137, and $N_{(Ela)}$ = 151, for a total of 420 sections. I use the variable *subject* as a potential source of information about the ambiguous fit specifying a set of subject-specific models for Model 5. This specification allowed me to evaluate model fit by subject and use the partially-saturated methodology to isolate the location of misfit.

Estimating separate, subject-specific models for the within and between levels (partially saturated method) revealed the source of ambiguous fit as the Algebra I sections at the between level ($RMSEA_{B2}$ = .131 and CFB_{B2} = .781; see Table 16. Not surprisingly and consistent with the findings across models in both years, the within levels of all three subjects also demonstrated consistent and acceptable levels of fit across all three subjects (Bio: $RMSEA_{W2}$ = .062 and CFB_{W2} = .995; ELA: $RMSEA_{W2}$ = .032 and CFB_{W2} = 997.; ALG = $RMSEA_{W2}$ = .057 and CFB_{W2} = .993. At the between level, Biology and English Language Arts revealed good levels of fit, (Bio - $RMSEA_{B2}$ = ..062 and CFB_{B2} = .995; ELA – $RMSEA_{B2}$ = .0553 and CFB_{B2} = .996. These results point to the Algebra I sections as the source of misfit for Model 5 in Year 2.

Table 15

Year 2 Test of Level-Specific Model Fit

Model 3	7-Factor								
Overall		25702.560	1146	0.00	.051	.771	.050	.048	
Classroom-level		5073.289	573	0.00	.137	.928	-	-	
Student-level		35848.444	573	0.00	.087	.970	-	-	
Model 5 RF	2-Factor								
Overall		818.359	50	0.00	.044	.981	.021	.011	
Classroom-level		118.961	25	0.00	.095	.981	-	-	
Student-level		1288.154	26	0.00	.079	.992	-	-	
Model 11	Wallace Bifactor								
Overall		14310.159	1164	0.00	.037	.877	.033	.052	
Classroom-level		5984.333	582	0.00	.149	.914			
Student-level		18161.755	582	0.00	.061	.985			
Model 12	Wallace Hybrid								
Overall		10950.921	1146	0.00	.032	.909	.029	.051	
Classroom-level		5722.238	573	0.00	.147	.918	-	-	11
Student-level		13428.525	573	0.00	.084	.972	-	-	12

Note. n_w = 8,120, n_b = 419; *RF = residual factor to account for cross loadings of the two reverse coded items (R_C15 with R_C27);); χ^2 = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; CFit = close fit probability that RMSEA < .05); CFI = comparative fit index; $SRMR_w$ = within level standardized root mean square residual; $SRMR_b$ = between level standardized root mean square residual.

Table 16

Year 2 Test of Level-Specific Model Fit by Subject

Models	Description	χ^2	df	*p*	RMSEA	CFI	SRMRw	SRMRb
Model 5 RF	All data							
Overall		818.359	50	0.00	.044	.981	.021	.011
Classroom-level		118.961	25	0.00	.095	.981	-	-
Student-level		1288.154	26	0.00	.079	.992	-	-
Model 5 RF	Biology							
Overall		358.618	50	0.00	.048	.977	.025	.010
Classroom-level		37.915	25	0.00	.062	.995		
Student-level		278.682	25	0.00	.0633	.995		
Model 5 RF	English Language Arts							
Overall		327.702	50	0.00	.042	.982	.022	.008
Classroom-level		28.897	25	0.00	.0322	.997		
Student-level		248.581	25	0.00	.0553	.996		
Model 5 RF	Algebra I							
Overall		228.468	50	0.00	.039	.986	.020	.024
Classroom-level		81.378	25	0.00	.131	.781		
Student-level		168.648	25	0.00	.0507	.993		

Note. Biology n_w = 2,671, n_b = 137; English n_w = 3,065, n_b = 150; Mathematics n_w = 2,351, n_b = 137; χ^2 = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; CFit = close fit (probability that RMSEA < .05); CFI = comparative fit index; $SRMR_w$ = within level standardized root mean square residual; $SRMR_b$ = between level standardized root mean square residual.

Year 2: Metric invariance. Based on the above results, I conclude this section by testing for metric invariance of Model 5, excluding the Algebra I observations from the analysis. As in the previous section, I fit four specifications of Model 5 (a–d) to systematically test for metric invariance. Specification 5a represented the equal factor structures across levels and resulted in strong model fit (χ^2 = 623.937 (50), RMSEA = .045, CFI = .980, and SRMRw = .022, and SRMRb = .008). In Specification 5b, I added the constraint of zero residual variation at the between level, which also fit the data well (χ^2 = 897.378 (59), RMSEA = .042, CFI = .980, and SRMRw = .021, and SRMRb = .026). I further constrain 5b to equal factor loadings across levels (Specification 5c) to represent strong metric invariance. This model also provided a high degree of fit to the data (χ^2 = 562.683 (66), RMSEA = .036, CFI = .981, and SRMRw = .022, and SRMRb = .056), albeit with some slippage in fit based on SRMRb. The final specification in this section is 5d, which frees the residual variances and also resulted in good fit overall with reduction in fit based on SRMRb (χ^2 = 587.193 (57), RMSEA = .040, CFI = .981, and SRMRw = .022, and SRMRb = .065). See Table 17 for the full results of the invariance testing.

These results from the Year 2 data provide additional evidence for metric invariance of Model 5 and thus valid comparisons of differences across classrooms based on the two-factor reduced form Tripod model, with the caveat that the Algebra I sections were excluded. In Table 18, I provide parameter estimates and standard errors for Model 5c (Year 2 data without the Algebra I sections).

Table 17

Year 2 Model 5 Test for Equality of Measurement Structure

Models	Description	χ^2	df	*p*	RMSEA	CFI	SRMRw	SRMRb
Two-Factor Reduced Form + Residual Factor								
Model 5a	Loadings + residuals estimated	623.937	50	0.0	.045	.980	.022	.008
Model 5b	Residuals @0	897.378	59	0.0	.042	.980	.021	.026
Model 5c	Equal loadings, residuals @0	562.683	66	0.0	.036	.982	.022	.056
Model 5d	Equal loadings, residuals estimated	587.193	57	0.0	.040	.981	.022	.065

Note. *n*w = 5,754, *n*b = 288χ^2 = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; CFit = close fit (probability that RMSEA < .05); CFI = comparative fit index; $SRMR_w$ = within level standardized root mean square residual; $SRMR_b$ = between level standardized root mean square residual

I also calculated the proportion of classroom-level variance for the two factors of Clarify and Captivate (.322 and .395, respectively). These values indicate a considerable degree of variance between classrooms and similar magnitudes to Year 1.

Table 18

Year 2 Parameter Estimates for the Reduced Form Tripod Model

	Within				Between			
	Est.	*SE*	*p*	Std	Est.	*SE*	*p*	Std.
Factor loadings								
Factor 1								
C11	1.141	.021	0.0	.752	1.141	.021	0.0	1.016
C12	1.360	.026	0.0	.806	1.360	.026	0.0	1.028
C13	1.305	.024	0.0	.794	1.305	.024	0.0	.971
C14	.898	.018	0.0	.668	.898	.018	0.0	1.058
R_C15	.593	.018	0.0	.464	.593	.018	0.0	.734
Factor 2								
C24	1.665	.029	0.0	.871	1.665	.029	0.0	.968
C25	1.561	.026	0.0	.841	1.561	.026	0.0	.996
C26	1.039	.018	0.0	.748	1.039	.018	0.0	.996
R_C27	.828	.017	0.0	.578	.828	.017	0.0	.960
Residual factor							-	
R_C15	1.00			.416	1.00		-	.510
R_C27	1.00			.376	1.00		-	.397
Factor Variances							-	
Factor 1	1.00	.00	999	1.00	.435	.00	0.0	1.00
Factor 2	1.00	.00	999	1.00	.594	.00	0.0	1.00
Residual	.282	.024	0.0	1.00	.070	.00	0.0	1.00
Factor covariance								
$Cov_{(F1, F2)}$	.874	.004	0.0	.874	.447	.039	.00	.939
$Cov_{(F1, RF)}$	.00	.00	999	-	.00	-	-	.00
$Cov_{(F2, RF)}$	.00	.00	999	-	.00	-	-	.00

Note. Nw = 5754; Nb = 288; Est. = factor loading estimate; *SE* = standard errors; *p* = probability level of .05; Std. = standardized estimate; *RF = residual factor to account for cross loadings of the two reverse coded items (R_C15 with R_C27).

In summary, the multilevel analysis of the Year 2 data provided consistent results to Year 1 across individual factors, as well as the single and multilevel models. As in Year 1, the two individual factors of Clarify and Captivate exhibited consistent fit and none of the hypothesized Tripod models met the requirement for equal factor structures across levels (configural invariance). In addition, the two bifactor models that showed promise at L1 in Year 1 also failed to achieve configural invariance between levels in Year 2. As in Year 1, only Model 5, achieved full isomorphic invariance (configural and metric invariance) in Year 2. However, an important caveat to this result is the exclusion of the Algebra sections from metric invariance testing.

Conclusion

This chapter presented a multilevel analysis of the Years 1 and 2 data for the Tripod Student Perception Survey. Consisting of five research questions, in each section I used a multilevel model evaluation strategy to determine the degree of isomorphism present the hypothesized factor models. The purpose of the analysis was to test the hypothesis that factor models designed to represent multiple dimensions of teacher effectiveness are valid measures of effective classroom instruction. I employed construct isomorphism as the standard by which to evaluate the validity claims of the Tripod Survey (Chan, 1998; Jebb et al., 2019; Ruelens, Meuleman, & Nicaise, 2018; Tay et al., 2014; Van Mierlo, Vermunt, & Rutte, 2009). I first evaluated the seven individual factors proposed by the Tripod authors as valid measures of effective teaching. Next, I evaluated 12 hypothesized combinations of those seven factors, suggested by the Tripod authors, Wallace et al. (2016) and other researchers within the teacher evaluation literature.

In RQ1 and RQ2, I determined model fit for the 12 proposed models at L1 based on model fit statistics. Three of the 12 models exhibited acceptable levels of model fit (Models 5, 11, and 12). Although Model 3 did not provide acceptable fit, I included it in subsequent analysis because of its widespread used in U.S. classrooms and to highlight its likely inappropriate use in evaluating teacher effectiveness. In RQ3, I determined that the degree of group-level variance, as evidenced by ICCs and the DEs, warranted a multilevel analysis. I next proceeded to fit a series of partially saturated models to evaluate fit for each of the four models the within and between levels. This process allowed me to determine the degree of configural invariance across Levels I and II of the analysis. Only Model 5, a reduced form of the Tripod model consisting of the factors of Clarify and Captivate, achieved configural invariance (i.e., configural isomorphism). Equal factor structure at both the student and classroom level is a prerequisite for metric invariance, which I then tested via RQ4. In RQ4, I evaluated Model 5 for metric invariance and determined, based on model fit, that the factor loadings were equivalent across levels (metric invariance). With metric invariance established, I concluded that differences in classroom-level variability could be attributed to differences across classrooms. The factors of Clarify and Captivate accounted for 27.1% and 32.7% of the differences, respectively in the nine items measuring these two dimensions of teacher effectiveness. The final research question RQ5, evaluated the same models from Year 1 on a new sample from the second year of the MET study. Here again, I found that only Model 5 achieved both configural and metric invariance, although with an additional caution that the Algebra I sections were excluded in order to achieve acceptable levels of

model fit. Excluding the Year 2 math sections provided comparable model fit to Year 1. Both Clarify and Captivate accounted for considerable differences between classrooms of 32.3% and 39.5%, respectively in Year 2. In other words, roughly 30 to 40% of variability in the items measuring these two dimensions of teacher effectiveness can be attributed to differences in individual classrooms and ultimately the teacher in each classroom.

Chapter 5: Discussion

The purpose of this quantitative study was to explore the viability of a high school student perception survey to serve as a valid and reliable indicator of teacher quality. As mentioned, little empirical evidence exists about the validity of these types of precollege perception surveys as aids in the teacher evaluation process (Geiger et al., 2019; Kuhfeld, 2017; Polikoff, 2017; Wallace et al., 2016). This lack of information has forced school leaders to make decisions about the appropriateness of using student perceptions in the evaluation of the teaching staff with limited information. To fill this information gap, I used a data set that included the high school version of the Tripod student perception survey, the most well-known and widely used precollege student perception survey in the United States (Gieger, 2019). I evaluated the psychometric properties of the Tripod with a multilevel factor analytic strategy at the student and classroom level and by combining two streams of analysis for evaluating configural and metric isomorphism. Multilevel isomorphism consists of two primary considerations across levels: (a) equal factor structure (configural) and (b) equal factor loadings (metric; see Table 6). With this approach, I contributed additional validity evidence to the teacher evaluation literature on precollege student surveys and highlighted a key methodological strategy for evaluating multilevel survey instruments. My study is the first to systematically evaluate the high school version of the Tripod survey for (a) the hypothesized factor structure using a partially saturated model methodology, (b) the degree to which the Tripod survey exhibits the prerequisite metric isomorphism needed for subsequent statistical analysis, and (c) replicability of results over a second sampling period.

The factor analysis yielded five key takeaways:

1. No empirical evidence was found in support of the hypothesized seven-factor structure of the Tripod survey as stated by the survey authors. Moreover, claims that the individual factors highlighted in the MET study as highly stable and reliable were not supported by the data at either the individual or classroom level.
2. No empirical evidence was found at the classroom level in support of alternative factor structures suggested by other researchers, including the higher-order bifactor models.
3. A two-factor model comprising the individual factors of Captivate and Clarify emerged as a potential measure of teacher effectiveness at both levels.
4. The reserve coded items, consistent with the findings from Wallace et al. (2016), were challenging for students to interpret.
5. The Year 1 results for the two-factor model and the reverse coded items were replicated in the Year 2 data, although only by excluding the algebra sections, which emerged as a potential source of bias.

Interpretation of the Findings

Conceptual Interpretations

Within the education literature, teacher quality has been shown to be an important component of students' academic, social, and emotional success with long-term consequences (Chetty et al., 2014; Rothstein, 2013). Though the economic and social value of highly effective teachers has been well-articulated, defining and measuring what

it means to be highly effective is much less well understood (Wallace et al., 2016). Further, the DMEE, which is an approach to explain the relationship between education and student outcomes with a focus on teacher quality, is contingent on the accurate assessment of teacher effectiveness. Although the MET study incorporated multiple measures of teacher effectiveness, the authors did not evaluate the psychometric properties of many of the instruments used in the study, including the Tripod (Jensen et al., 2018). The study authors assumed that the Tripod structure exhibited construct validity (Kane et al., 2013). Without basic information about the validity and reliability of the Tripod, it is difficult to make meaningful interpretations of the MET findings. Additionally, few peer-reviewed studies have investigated the construct validity of the Tripod, and none have investigated the similarity of structure across levels (within and between classrooms) necessary for aggregation and comparison between classrooms (Jak, 2019). Understanding these factor structures are a basic perquisite for establishing validity arguments for the use of the Tripod in the evaluation of teachers (Jak, 2013, 2014, 2019; Stapleton 2016).

Addressing these knowledge gaps regarding teacher effectiveness and the Tripod survey, my study points to several key findings. First, following the well-documented multilevel strategy for model evaluation, I began by evaluating the construct of teacher effectiveness at the individual level. This estimation ignores the multilevel nature of the data in order to evaluate the basic structure of the construct (Dyer, 2005; Muthen, 1994). I tested each of the seven individual factors followed by 12 different specifications of the Tripod and found that only two of the seven individual factors, Clarify and Captivate,

provided consistently good fit across the fit indices. These two factors highlight what might be characterized as cognitive activation, as the items assess the teacher's ability to explain concepts, work toward student understanding, capture interest, and keep students meaningfully engaged. Further, these qualities of teaching may be more easily perceived through the eyes of students compared to the other Tripod factors in that students who experience these qualities can identify and rate them. Other factors may be clear in the eyes of practiced observers, but the Tripod items did not allow students to distinguish those qualities.

Second, in my analysis, the factor of control, which is most often referred to in the MET study and by other authors as the strongest indicator of effective teaching, demonstrated the poorest fit of all the individual factors. This inconsistent finding also points to potential problems with rating effective teaching through student observation. Because my study is the first to evaluate the factor structure of the Tripod using the Grade 9 data, younger students studied in previous research may have perceived classroom qualities differently because they were less mature than these older students. In addition, the selection of the type of statistical approach in modeling multilevel data may also account for the difference in my findings on the factor of control.

Along with the individual factors, I also tested a series of multifactor models, based on the Tripod configuration, at the student level (see Figures 4–6.) None of the hypothesized Tripod models met acceptable levels of fit across multiple model fit indices. This result included the well-known seven-factor model as well as the higher-order model the Tripod is named after. However, my results were consistent with other authors who

found no evidence of a seven-factor model in their analyses with younger students in earlier grades (Kuhfeld, 2016, 2017; Wallace, 2016). Though the Tripod was designed to capture teacher quality across seven dimensions, the structure was not evident in my analysis nor in analyses by other researchers. These results raise questions about the validity of the survey and the ability of students to evaluate their teacher's performance. Further, I found no evidence to support a higher-order structure represented by three broader categories of instructional effectiveness: (a) personal support, (b) curriculum support, and (c) academic press, as designed to account for variation in the seven first-order factors. This finding suggests that the Tripod, as hypothesized, does not support a meaningful interpretation of classroom differences in teacher performance at the high school level. These results indicate, from my analysis and others' the inadequacy of the 36 Tripod items, as rated by students, to measure seven dimensions of effective teaching (Geiger, 2019).

I also estimated a number of bifactor models, including a specific model developed by Wallace et al. (2016). Their research, based on the middle school math data collect as part of the MET study, also showed no evidence of the Tripod's seven-factor structure. However, they tested several alternative configurations, including the idea that a general construct of effective teaching might better represent the data while also maintaining several distinct but ancillary subfactors. This type of configuration, as shown in Figure 7, is a special form of the higher-order factor structures referred to in the literature as a bifactor model (see Figure 10). Wallace et al. did find evidence for this general teaching factor along with two subfactors: the Control factor (i.e., classroom

management) and a new method factor that included all the negatively worded items in the Tripod (includes five items). Wallace et al. found this model to fit the data fairly well. Using the Grade 9 data, I replicated their results at the student level and found that a general factor and two subfactors provided adequate fit to the data. The bifactor approach that Wallace et al. deployed was insightful in that a general factor accounted for the variation in the Tripod items rather than seven covarying factors. Additionally, Wallace et al. suggested the negatively worded items might be difficult for students to interpret and serve as a method factor in the model. My results confirmed these two points for the Grade 9 data at the individual level of the analysis.

In addition to the multifactor and bifactor models, I also fit a new two-factor model based on my findings that only two individual factors, Clarify and Captivate, emerged as strong fit to the data (see Figure 11). Further, because both factors also included a negatively worded item, I also added a third subfactor based on those two negative items to account for same method factor as specified by Wallace et al. (2016), and as a way to further explore this potential method issue (Jak, 2019). The model, which I refer to this as the reduced form of the Tripod, which provides the best and most consistent fit to the data of all the models evaluated at L1. Together, the factors of Clarify and Captivate may appear to students as something that engages their learning process and is more readily recognized as effective teaching. To summarize my findings at the individual level, the hypothesized Tripod structures were not supported by the data; however, two bifactor alternatives and a new reduced form version of the Tripod did show promise at this step of the validation process.

I next discuss the findings from the multilevel perspective, where I evaluated the extent to which the construct of teacher effectiveness as defined by structures at L1 hold at L2 (i.e., isomorphic between students and classrooms). If the factor structures are different across levels—in other words, L1 structure differs from L2—the constructs have a different meaning across levels (Jak 2013, 2019; Rabe-Hesketh, Skrondal, & Pickles, 2004; Rhu & West, 2009; Tay et al., 2014). Without consistent meaning across levels, L1 student ratings cannot be aggregated to represent L2 classroom characteristics for evaluation of individual teachers. However, if the structures are isomorphic, then aggregation of L1 ratings to form L2 measures is a valid process. Thus, it is imperative to assess the degree of similarity between levels based on the structures verified at L1. Accordingly, I brought forward the three best-fitting models from the single level analysis: (a) the two-factor reduced form of the Tripod (Clarify and Captivate, Model 5); (b) the Wallace bifactor (Model 11); and (c) a hybrid of the Wallace model, where I also included Clarify and Captivate as subfactors along with Control and the method factor (see Figure 12). In addition, I also included the seven-factor Tripod model in this analysis because, although the fit was questionable at Level 1, its use in thousands of schools across the United States is based on little empirical research, and the results may be useful to school officials (Geiger, 2019). Thus, adding it into this phase of the analysis provided needed information not currently available about the psychometric properties of the Tripod.

The key consideration in the construct validation process is the degree of isomorphism across levels. As stated earlier, it is important to understand the nature and

relationship between constructs that exist at different levels; otherwise, it is unclear what L2 construct represents (Hox et al., 2017). In the case of teacher evaluation, student ratings are collected at L1. The aggregated measures are then used to evaluate L2 classrooms. If the factor structure does not hold between levels, it calls into question what aspects of teacher effectiveness is represented at L2. Psychometric isomorphism provides a framework in which to assess differences in factor structures across levels (see Table 6) and to evaluate if the aggregation of L1 data is appropriate for measurement of an L2 phenomenon.

I tested isomorphism in a two-step process by determining (a) if the number of factors and the pattern of loadings at L1 were similar at L2 (i.e., configural isomorphism) and (b) if the magnitude and size of the factor loadings at L1 are mirrored at L2 (i.e., metric isomorphism; Jak, 2013; Jebb et al., 2019; Tay et al., 2014). In my examination of configural isomorphism, I fit three models, an overall multilevel model and two partially-saturated models. One focused on the within-level and the other focused on the between level (see Table 2). In comparison to the overall model fit, I found that three of four models exhibited stronger fit at L1 and exhibited substantial decrement in fit at L2. These findings help explain the ambiguous fit of models from the single-level analysis. Essentially the seven-factor Tripod and the two bifactor models did not exhibit the same factor structures between L1 and L2 and thus are not configurally isomorphic—that is, the factor structures at L1 do not have a corresponding structure at L2. The exception to these results was found in Model 5, the two-factor reduced form Tripod model, which had excellent fit to the data at both L1 and L2. This finding provides evidence that equal

factor structures exist across levels; thus, configural isomorphism holds for Model 5. This implies that the construct meaning for Clarify and Captivate the factors in Model 5 have the same meaning across levels. These results also suggest that the seven-factor model and the bifactor models more closely fit the data at L1, but that those same structures did not hold at L2. Thus, the construct of teacher effectiveness as defined by the seven-factor Tripod and the alternative bifactor models have different structures across levels and thus different construct meaning across levels.

With only Model 5 of the four models meeting the requirement for configural isomorphism, I then evaluated Model 5 for metric isomorphism. This step determines the degree to which the factor loadings are metrically equivalent across levels and is a necessary requirement for valid comparisons of variation across groups (Jak, 2014, 2019; Rabe-Hesketh et al., 2004). Without metric isomorphism, comparisons of teacher effectiveness across classrooms are not warranted and raise concerns about what construct is being measured at L2 (Tay et al., 2014). My evaluation of Model 5 resulted in strong metric isomorphism. The items measuring the factors of Clarify and Captivate maintained equal factor loading across L1 and L2. This finding provided evidence that the variance between classrooms on the teacher qualities of Clarify and Captivate are psychometrically valid for comparisons. Based on this result, I calculated the L2 variation across classrooms to be 27.1% for the factor of Clarify and 32.7% for Captivate. This is interpreted as the amount of total variance in the items that can be accounted for by each factor at the classroom level, such that roughly 30% of the variance in these factors is attributable to the classroom level and by extension to the individual teachers. The

remaining 70% of the total variance is then, by definition, accounted for at the student level. These findings contribute important psychometric information about the validity of the constructs hypothesized to measure teacher effectiveness.

In summary, the two-factor reduced form of the Tripod exhibited full psychometric isomorphism (e.g., configural and metric). The degree to which the other models were not configurally equivalent suggests different factor structures across levels. To the extent that the Tripod and bifactor model exhibited stronger fit at the individual rather than classroom level, student perceptions of their classrooms referenced more individual views of students rather than indexing a more collective point of view. Students' perceptions may have reflected more of an idiosyncratic difference than a shared perspective about their classroom teacher. Further, even with self-reference as a given, it was somewhat surprising to find less consensus among students within the same classroom on more of the items. Clearly, other factors are at play in the classroom that were not captured by the 36 Tripod items (Jak. 2013).

Importantly, the negatively worded items also seemed to contribute to an effect not intended by the Tripod authors. Of the 36 item, five items are negatively worded and served the purpose of disrupting the positive direction of the response pattern. Unfortunately, these items seemed to have confused students. These items generally had the lowest factor loadings of all the items in any of the models. Further, I tested a residual factor to account for this method issue in Model 5, the reduced-form of the Tripod, and it provided stronger fit to the data than the model without the residual factor. Wallace et al.

(2016) also accounted for these items by including a factor with all five of the negatively worded items from the full Tripod model and found this factor to be significant as well.

The final result to highlight is the replication of findings from Y1 to Y2. The design of the MET study provided an excellent opportunity to evaluate the strength of the Y1 findings on a second year of data. My Y1 findings were essentially replicated in Y2. I evaluated each of the seven individual factors as well as the 12 models tested in Y1. I was able to replicate finding at each step in the multilevel validation process, including (a) identifying the same three single level models to bring into the multilevel analysis, (b) the poor fit of all but Model 5 at L2, and (c) the relative strength of the models at L1. As in Y1, student perceptions did not coalesce at the group level for three out of the four models (Models 3, 11, and 12). Model 5, the best and most consistently well-fitting model in Y1, however, yielded a substantial decrement fit over Y1, at both L1 and L2. This reduction in fit was perplexing, given the relatively consistent findings for the other models between years. To explore a possible explanation, I exploited a variable in the data set that described the academic subject being taught in each section to help explain the inconsistent fit of Model 5. Although not ideal, I fit each of the topic areas separately (Algebra I, biology, and English language arts) and determined that the reduction in fit was isolated in the Algebra I sections. Once I removed those sections for the Model 5 analysis, the model performance returned to Y1 levels. The poor fit of the Algebra I sections was located at L2 and may be related to the random assignment of teachers to classrooms. This random assignment was the case for all subjects, but students might be more knowledgeable and sensitive to teacher reputations and teaching styles for

mathematics teachers more generally. Having thought students were sorting into one type of instructional style (classroom) and then finding themselves in another that potentially did not align with their instructional needs may have resulted in a bias in those ratings. The effects of student sorting are well known in terms of introducing bias into the analysis (Rothstein, 2009), and these results seem to support that finding as well.

Methodological

Finally, the methodological findings are important to note. This study was the first to utilize psychometric isomorphism as a multilevel construct validation strategy with the MET data generally and the Tripod items specifically. Further, isomorphism is a key requirement for establishing measurement validity for an instrument and a necessary step before use of the instrument in downstream analysis such as incorporating direct effects in a structural model (Jak 2013, 2019; Rabe-Hesketh et al., 2004). Using psychometric isomorphism in the multilevel educational psychology literature is unusual and seen more often in the organizational psychology literature on the study of cultures (Ruelens, 2018). My study utilized the construct validation framework outlined by Jebb et al., (2019) and Tay et al., (2014), while bringing together two separate strategies for (a) configural isomorphism by testing specific levels of fit/misfit using partially-saturated models (Ryu & West, 2009), and (b) metric isomorphism by testing the equality of factor structures with zero residual variances (Jak, 2013). This synthesis of approaches allowed me to extend the work of Wallace et al. (2016) as well as other researchers who have evaluated the Tripod survey. By systematically identifying the specific level of model misfit and thus the degree of isomorphism present in the data, I more fully assessed the multilevel

factor structure of the Tripod survey. This level of detail contributes both to the validity evidence regarding the Tripod specifically and contributes to the process of describing and analyzing multilevel construct validation more generally.

Limitations

Limitations in the study can be grouped into three categories: (a) MET study design, (b) the instrument, and (c) data and analysis. In terms of the design, the MET study is the largest and most ambitious study of teacher effectiveness ever undertaken in the United States (Jensen et al., 2019). The sample size was more than adequate for multilevel analysis, particularly at the group level. Most methodologists agree group sample sizes of 100 are considered large and the MET group samples used in my analysis were 1,188 and 419, in Y1 and Y2, respectively. While the size of the study was impressive, the convenience sampling and the attrition of teachers between Y1 and Y2 hindered the generalizability of results (Jensen et al., 2019). In addition, the design of the study did not include a standardized test for the Grade 9 students, which limited the ability to validate findings against an outside criterion, such as a standardized test.

Second, the Tripod instrument documentation was lacking basic information about the theoretical framework as well as the psychometric properties of the survey (M. Kane et al., 2013). In addition, one of the factors, Care, had only three indicators, making it impossible to evaluate its structure. Also of note, the individual items had inconsistent reference points, with some items asking students to rate their classroom, while other items referenced the students' point of view (e.g., see Item 7 versus Item 8 in Table 4). This inconsistency complicated the interpretation of student perceptions and potentially

contributed to the lack of isomorphism exhibited by the full seven-factor model. As discussed in Chapter 2, construct isomorphism depends on the hypothesized level and composition model employed for the construct (Mehta & Neale, 2005; Tay et al., 2014). The Tripod authors did not explicitly state where, along the continuum of composition models, the Tripod was to be positioned. In addition, the negatively worded items, as discussed previously, added another complication to the assessment of teacher effectiveness as students appeared to have difficultly interpreting the meaning of those items.

In addition, the use of student perceptions as a measure of teacher effectiveness is relatively new at the pre-college level and not without controversy (Geiger, 2019; Marsh, 2019). As a staple in the postsecondary landscape, students are generally accepted as appropriate raters of their college instructors (Marsh, 2019). However, a recent systematic review of 28 postsecondary studies on the use of student course evaluations identified several sources of student bias including instructor characteristics, grading leniency, and course load (Wang & Williamson, 2020). These issues call into question the validity of student raters at the postsecondary level and, by extension, encourage the cautious use of student ratings at the pre-college level as well.

The final area of limitations to discuss is the analysis. The Tripod data are categorical and measured on a 5-point scale. This type of data calls for a particular type of estimator that is consistent with non-normal data. The WLSMV estimator has been found to be less biased and more accurate than maximum likelihood estimators for categorical data (Li, 2016), even when a non-normal correction is added (i.e., robust

maximum likelihood estimator). However, much of the literature on partially saturated model estimation used robust maximum likelihood estimators, which presume interval scale measurement, even when the data were clearly categorical. In order to overcome some of the limitations associated with estimating the independence model with WLSMV (MPlus discussion board, per Muthen, 2009), I conducted all the partially saturated model analysis using the WLSM but for interpreting parameters, I utilized WLSMV, again per Muthen (2009). I also tested the WLSM results against robust maximum likelihood and found I was able to mostly replicate the results. In addition, the clear misfit of models at L2 suggested that the structures were different between levels. Investigating specific alternative structures at L2 is a logical extension of the current study but beyond the scope of this work. Also of note, the use of language to describe construct validation of multilevel models can be challenging. I adopted the term isomorphism to describe cross-level similarities between constructs. Other researchers have used terms such as *equivalence* or *invariance* to describe the same pattern of equal factor structures and loadings across levels or the violation of those patterns and loadings. This inconsistent use of language to describe essentially the same phenomenon is confusing and somewhat problematic for unifying the process of multilevel construct validation.

Recommendations

Clearly the work in multilevel construct validation is evolving (Tay et al., 2014). Combining (a) partially saturated methods (Rhy & West, 2009) for determining configural isomorphism and (b) evaluating the equality of loadings (Jak, 2013, 2014) to establish metric isomorphism can provide a more complete picture of psychometric

isomorphism, a necessary condition for validating group measures based on the aggregation of individual ratings. Further, extending the analysis of psychometric isomorphism to the elementary- and middle-school grades would also aid in clarifying the structure of the Tripod and determining more precisely where model misfit may occur. Also of interest is the exploration of the factor structure at L2. Given the L2 structure in all models, with the exception of the two-factor reduced form of the Tripod, did not correspond to the structure at L1 it would be worthwhile to investigate the structure at L2.

Another area for future research is the use of item response theory to more carefully assess the functioning of the individual items. My study focused on the structure of teacher effectiveness at the individual and group level using factor analysis methodology. A natural extension would be to explore item characteristics using item response theory models. These types of models allow for the investigation of individual item discrimination and difficulty, which are useful characteristics in determining the quality of survey items. Given the lack of support for a seven-factor model, it would be useful to explore more thoroughly the individual items of the Tripod.

A final area of meaningful exploration is the use of reverse-coded items. While the general benefit of minimizing a pattern of response is usually useful, it appeared in this study that students were confused about the negatively worded items. Careful evaluation and testing of negatively worded items before their use in surveys of student perceptions is warranted.

Social Change Implications

Of all the resources a school can directly control, the classroom teacher matters most for student outcomes (Opper, 2019). Exposure to a high-quality teacher is associated with critical student outcomes including improved college attendance and higher salaries (Chetty et al., 2014). Clearly, educational opportunity plays a pivotal role in improving the life chances of children (Blanden, 2020) and has implications for positive social change. Ensuring that every child has the opportunity to be taught by an effective teacher is imperative if we are going to improve outcomes for individuals as well as for society more broadly. Measuring teacher performance is a prerequisite to ensuring each student has a quality teacher. As discussed in this study, evaluating teacher quality has proven to be challenging, particularly in the use of student surveys in the evaluation process. That is not to say that conceptually, pre-college student surveys are inherently flawed; rather, additional work is needed to develop instruments that are appropriately measuring the construct of teacher quality. The survey investigated here, while conceptually attractive, failed to provide empirical evidence of construct validity at the classroom level for all configurations tested, save one. Without evidence that the items measure teacher effectiveness, use of the Tripod to evaluate dimensions of teacher performance appears premature and warrants additional investigation. As documented in a 2019 report on the state of pre-college student perception survey usage, there is no peer-reviewed research to support the widespread use of these types of surveys for teacher evaluation (Geiger, 2019). Further, the report advised users to take care in making determinations about teacher effectiveness with instruments that are potentially unreliable

and invalid (Geiger, 2019). My results highlight the problematic nature of measuring teacher effectiveness with a tool that does not generalize from the student to the classroom level. That said, the MET data set is just one instance of the reported use of the Tripod, and additional data would need to be collected to augment the validity evidence presented here. My study is thus cautionary: Although I applaud the attempt to incorporate the views of students in the evaluation process, these results suggest additional evidence is needed to support widespread adoption of pre-college surveys.

The use of multilevel construct validation tools is a critical component in assessing the viability of survey instruments such as the Tripod. In a perfect world, using these tools would allow researchers to pinpoint areas of misfit in evaluating constructs that are hypothesized to exist concurrently across levels or as separate and different constructs. Previous work evaluating the Tripod either assumed equal factor structures or did not explore the possibility of different structures between levels. Evaluating psychometric isomorphism can help identify the appropriateness of survey instruments and provide missing analysis about the validity of the Tripod and other pre-college survey instruments. Without precise information about model fit, factor structures are difficult to evaluate in a systematic fashion. Furthermore, as noted previously, establishing the basic psychometric properties of these type of instruments is a necessary condition for making subsequent evaluations of teacher practice and effectiveness.

A final area of importance is the use of psychometric isomorphism (configural and metric) in the discussion of composition models. As highlighted by Tay et al. (2014) and others, the aggregation process to generate scores requires a hypothesis about how

the measurement at one level reflects the construct at another level (Bonito & Keyton, 2019; Chan, 1998; Jebb et al., 2019). The way in which a total score is produced or aggregated depends on the type of composition model, which the Tripod authors did not explicitly state. As noted in Chapters 2 and 3, for the purposes of my study and based on my assessment that the majority of the items attempted to reference the collective construct of teacher effectiveness, I assumed a reference-shift model for the aggregation process. Referent-shift models are measured at the individual level with the average scores representing the shared perception of each group on a particular construct (Jebb et al., 2019). Psychometric isomorphism tests the extent to which the assumption of a shared perception is valid. As my results indicated, there was little evidence to support shared perceptions at L2 because of the lack of isomorphism in all of the models except Model 5. Testing the composition model is accomplished by assessing the degree of isomorphism through the use of multilevel modelling tools (partially saturated and measurement invariant modeling). Future investigations in evaluating the Tripod and other pre-college surveys should at least consider the isomorphic properties of the instrument under investigation.

Conclusions

Teaching is complex, nuanced, and perplexingly difficult to measure. We have all experienced great teaching, as well as its opposite. But what is it that makes a great teacher? Many theories coalesce around a similar set of factors designed to represent what we think we all recognize as effective instruction. These factors, while intuitive, have proven to be extremely difficult to operationalize. The Tripod survey attempted to

define effective teaching into three broad dimensions that were further hypothesized to explain variation in seven distinct factors. As a theory of effective teaching, these seven factors purported to represent qualities of effective teaching that could be assessed through the eyes of pre-college students. While this theory makes logical and intuitive sense, the empirical evidence suggests that the 36 items that compose the Tripod instrument do not capture effective teaching as hypothesized. Or alternatively, students were unable, via the Tripod items, to distinguish the hypothesized factors. Rather, the Tripod seven-factor model and its variants appeared to not index a shared perspective from students that could be generalized to classroom teachers. Without supporting validity evidence, the theory of effective instruction, as operationalized by the 36 items of the Tripod, does not appear to be supported in the MET data for Grade 9 students. Further, several other models were also not supported, including a bifactor specification that had previously been reported as a potential alternative (Wallace, 2016). Interestingly, a two-factor reduced-form of the Tripod appeared to hold some promise of explain teacher effectiveness through the perspective of student rater.

All of these models were evaluated using a complex process that explicitly examined construct validity using a toolkit of multilevel methods. These methods were necessary to investigate the factor structure of the proposed models and also demonstrated a multilevel application of the toolkit on an important data set. The ability to identify appropriate measures of teaching effectiveness is a critical step in distinguishing and improving classroom teaching. Much research shows that teachers matter in the instructional process. Education leaders need instruments that are equally

clear in evaluating the quality of teaching. Because the Tripod does not appear to meet minimum requirements as a valid measurement instrument of the seven constructs it purports to measure, further work needs to be done. Every student deserves a high-quality teacher, and every teacher deserves to be validly assessed.

References

Aldeman, C. (2017). The teacher evaluation revamp, in hindsight. *Education Next, 17*(2). 60–68. Retrieved from https://www.educationnext.org/the-teacher-evaluation-revamp-in-hindsight-obama-administration-reform/

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

Balch, R. T. (2012). *The validation of a student survey on teacher practice* (Doctoral dissertation). Available from https://www.semanticscholar.org/paper/The-validation-of-a-student-survey-on-teacher-Balch/517b01f26e8c5c5bb57e0c66d16450c4d78fd70e

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. doi:10.1080/10627197.2012.715014

Berg-Jacobson, A. (2016). *Teacher effectiveness in the Every Student Succeeds Act: A discussion guide.* Retrieved from https://www.air.org/sites/default/files/downloads/report/TeacherEffectiveness_ESSA.pdf

Bill and Melinda Gates Foundation. (2018). *Measures of effective teaching: 1–Study information*. doi:10.3886/ICPSR34771.v3

Birman, B., Le Floch, K. C., Klekotka, A., Ludwig, M., Taylor, J., Walters, K., Garet, M.

S. (2007). *State and Local Implementation of the No Child Left Behind Act: Volume II-Teacher Quality under NCLB: Interim Report.* Washington, DC: U.S. Department of Education.

Birman, B., Boyle, A., Le Floch, K., Elledge, A., Holtzman, D., Song, M., & Yoon, K. (2009). *State and local implementation of the No Child Left Behind Act: Volume VIII—Teacher quality under NCLB: Final report*. Washington, DC: U.S. Department of Education.

Bollen, K. (1989). Structural equations with latent variables. doi:10.9781118619179

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145*, 27–41. doi:10.1016/j.jpubeco.2016.11.006

Brown, T. A. (2014). Confirmatory factor analysis for applied research. New York, NY: Guilford.

Byrne, B. M., & van de Vijver, F. J. (2014). Factorial structure of the family values scale from a multilevel-multicultural perspective. *International Journal of Testing, 14*(2), 168–192. doi:10.1080/15305058.2013.870903

Campbell, S. & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for*?* American Educational Research Journal, 55(6), 1233-1267. doi:10.3102/0002831218776216

Cavanagh, S. (2014, October 7). Company known for student surveys unveils new tool for teachers. *EDWeek*. Retrieved from https://marketbrief.edweek.org

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of*

Applied Psychology, 83(2), 234–246. doi:10.1037/0021-9010.83.2.234

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13. doi:10.1111/j.1745-3992.2009.00165.x

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh public schools. *Regional Educational Laboratory Mid-Atlantic*. Retrieved from https://files.eric.ed.gov/fulltext/ED545232.pdf

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12*(3), 471–492. doi:10.1207/s15328007sem1203_7

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593–2632. doi:10.1257/aer.104.9.2593

Cizek, G. J. (2016). Progress on validity: the glass half full, the work half done. *Assessment in Education: Principles, Policy & Practice, 23*(2), 304–308. doi:10.1080/0969594x.2016.1156642

Cohen, J., & Goldhaber, D. (2016a). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378–387. doi:10.3102/0013189x16659442

Cohen, J., & Goldhaber, D. (2016b). Observations on evaluating teacher performance: Assessing the strengths and weaknesses of classroom observations and value-

added measures. In J. A. Grissom & P. Youngs (Eds.), I*mproving teacher evaluation systems* (pp. 8–21). New York, NY: Teachers College, Columbia University.

Creemers, B. (2002). *The comprehensive model of educational effectiveness: Background, major assumptions and description.* Retrieved from: https://www.rug.nl/staff/b.p.m.creemers/the%20comprehensive%20model%20of%20educational%20effectiveness.pdf

Creemers, B., & Kyriakides, L. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement, 17*(3), 347–366. doi:10.1080/09243450600697242

Creemers, B., Kyriakides, L., & Antoniou, P. (2013). Establishing theoretical frameworks to describe teacher effectiveness. In *Teacher professional development for improving quality of teaching* (pp. 101–135). Dordrecht, Netherlands: Springer.

Danielson, C. (2011). Enhancing professional practice: A framework for teaching. Alexandria, VA: ASCD.

Danielson, C. (2012). Observing classroom practice. *Educational Leadership, 70*(3), 32–37. Retrieved from http://www.ascd.org/publications/educational-leadership/nov12/vol70/num03/Observing-Classroom-Practice.aspx

de Lima, J. Á., & Silva, M. J. T. (2018). Resistance to classroom observation in the context of teacher evaluation: teachers' and department heads' experiences and perspectives. *Educational Assessment, Evaluation and Accountability, 30*(1), 7–

26. doi:10.1007/s11092-017-9261-5

Dee, T. S., & Jacob, B. A. (2010). The impact of no child left behind on students, teachers, and schools. *Brookings Papers on Economic Activity, 2010*(2), 149–194. doi:10.1353/eca.2010.0014

Doan, S., Schweig, J. D., & Mihaly, K. (2019). The consistency of composite ratings of teacher effectiveness: evidence from New Mexico. *American Educational Research Journal*, *56*(6), 2116–2146. doi:10.3102/0002831219841369

Doherty, K. M., & Jacobs, S. (2015). *State of the states 2015: Evaluating teaching, leading, and learning.* Washington, DC: National Council on Teacher Quality.

Doyle, W. (1977). 4: Paradigms for eesearch on teacher effectiveness. *Review of Research in Education, 5*(1), 163–198. doi:10.2307/1167174

Dunn, E. C., Masyn, K. E., Johnston, W. R., & Subramanian, S. (2015). Modeling contextual effects using individual-level data and without aggregation: an illustration of multilevel factor analysis (MLFA) with collective efficacy. *Population Health Metrics, 13*(1), 12. doi:10.1186/s12963-015-0045-1

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *Leadership Quarterly, 16*(1), 149–167. doi:10.1016/j.leaqua.2004.09.009

English, D., Burniske, J., Meibaum, D., & Lachlan-Haché, L. (2015). *Uncommon Measures: Student surveys and their use in measuring teaching effectiveness*. Retrieved from http://www.air.org/sites/default/files/Uncommon-Measures-Student-Surveys-Guidance-Nov-2015.pdf

Fauth, B., Wagner, W., Bertram, C., Göllner, R., Roloff, J., Lüdtke, O., . . . Trautwein, U. (2019). Don't blame the teacher? The need to account for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology*. doi: 10.1037/edu0000416

Ferguson, R. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24–28. doi: 10.1177/2F003172171209400306

Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and Tripod 7Cs evidence distinguish key components of effective teaching. In T. Kane, K. Kerr, & R. Pianta (Eds.), Designing teacher evaluation systems: New guidance from the measures of effective teaching project (pp. 98–43). San Francisco, CA: Jossey-Bass.

Ferguson, R., & Ramsdell, R. (2011). *Tripod classroom-level student perceptions as measures of teaching effectiveness*. Paper presented at the Cambridge, MA: The Tripod Project. Retrieved from https://www.nnstoy.org/download/evalution/NCTE%20Conference%20Tripod.pdf

Finch, H., & Bolin, J. (2017). *Multilevel modeling using Mplus.* Boca Raton, FL: CRC Press.

Fleenor, L. (2015). *The relationship between student perceptions of classroom climate and TVAAS Student achievement scores in Title I schools.* Retrieved from https://www.semanticscholar.org/paper/The-Relationship-Between-Student-Perceptions-of-and-Fleenor/7209ea6b6ce3744fc036ece7d5f0349d5d4bb235

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466–491. doi:10.1037/2F1082-989x.9.4.466

Geiger, T., & Amrein-Beardsley, A. (2019). Student perception surveys for K-12 teacher evaluation in the United States: A survey of surveys. *Cogent Education, 6*(1). doi: 10.1080/2331186x.2019.1602943

Glaser, D. N. (2002). Structural equation modeling texts: A primer for the beginner. *Journal of Clinical Child & Adolescent Psychology, 31*(4), 573–578. doi:10.1207/S15374424JCCP3104_16

Goe, L., Bell, C., & Little, O. (2008). Approaches to evaluating teacher effectiveness: A research synthesis. *National Comprehensive Center for Teacher Quality*. Retrieved from https://www.wested.org/wp-content/uploads/goe-research-synthesis.pdf

Goos, M., & Salomons, A. (2016). Measuring teaching quality in higher education: assessing selection bias in course evaluations. *Research in Higher Education*, 1-24. doi:10.1007/s11162-016-9429-8

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), 444–470. doi:10.1086/2F669901

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*(1), 5–28. doi:10.1007/2Fs11092-013-9179-5

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*(3), 466–479. doi:10.1016/2Fj.econedurev.2010.12.006

Hanushek, E. A. (2016). School human capital and teacher salary policies. *Journal of Professional Capital and Community, 1*(1), 23–40. doi:10.1108/2Fjpcc-07-2015-0002

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal, 51*(1), 73–112. doi:10.3102/0002831213517130

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7–8), 798–812. doi:10.1016/j.jpubeco.2010.11.009

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record, 116*(1), 1–28. Retrieved from https://scholar.harvard.edu/mkraft/publications/state-and-local-efforts-investigate-validity-and-reliability-scores-teacher

Heck, R. H., & Thomas, S. L. (2015). An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus. Abington, England: Routledge.

Hess, F. M., & Eden, M. (2017). The Every Student Succeeds Act: What it means for

schools, systems, and states. Cambridge, MA: Harvard Education Press.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., 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. doi:10.1080/07370000802177235

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794–831. doi:10.3102/0002831210387916

Hinchey, P. H. (2010). Getting teacher assessment right: What policymakers can learn from research. Boulder, CO: National Education Policy Center. Retrieved from https://nepc.colorado.edu/publication/getting-teacher-assessment-right

Hox, J. (2010). Multilevel analysis: Techniques and applications. New York, NY: Routledge.

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. New York, NY: Guilford Press.

Huber, S. G., & Skedsmo, G. (2016). Teacher evaluation—accountability and improving teaching practices. *Educational Assessment, Evaluation and Accountability, 2*(28), 105–109. doi:10.1007/s11092-016-9241-1

Hull, J. (2013). Trends in teacher evaluation: How states are measuring teacher performance. *Center for Public Education*. Retrieved from https://gtlcenter.org/products-resources/trends-teacher-evaluation-how-state-are-measuring-teacher-performance

Jak, S. (2016). Testing and explaining differences in common and residual factors across many countries. *Journal of Cross-Cultural Psychology, 48*(1), 75–92. doi:10.1177/0022022116674599

Jak, S. (2019). Cross-level invariance in multilevel factor models. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(4), 607–622. doi:10.1080/10705511.2018.1534205

Jebb, A. T., Tay, L., Ng, V., & Woo, S. (2019). Construct validation in multilevel studies. In S. E. Humphrey & J. M. LeBreton (Eds.), *The handbook of multilevel theory, measurement, and analysis* (pp. 253–278). American Psychological Association. doi.org/10.1037/0000115-012.

Jennings, J. F. (2001). *Title I: Its legislative history and its promise*. In G. D. Borman, S. C. Stringfield, & R. E. Slavin (Eds.), *Title I: Compensatory education at the crossroads* (pp. 1–23). Mahwah, NJ: Erlbaum.

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

Jones, L. V., & Thissen, D. (2006). A history and overview of psychometrics. *Handbook of Statistics, 26*, 1–27.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. doi:10.1037/0033-2909.112.3.527

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17–64). Westport, CT: Praeger.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review, 42*(4), 448–457. doi:10.1080/02796015.2013.12087465

Kane, M. (2016). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy & Practice, 23*(2), 309–311. doi:10.1080/0969594x.2016.1156645

Kane, T. J., & Staiger, D. O. (2010). Learning about teaching: Initial findings from the Measures of Effective Teaching Project. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from https://docs.gatesfoundation.org/documents/preliminary-findings-research-paper.pdf

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

Kane, T. J., McCaffrey, D. M., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from https://k12education.gatesfoundation.org/resource/have-we-identified-effective-teachers-validating-measures-of-effective-teaching-using-random-assignment/

Kane, T., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Research Paper. MET Project. Seattle, WA. *Bill & Melinda Gates Foundation*. Retrieved from https://k12education.gatesfoundation.org/resource/gathering-feedback-on-teaching-combining-high-quality-observations-with-student-surveys-and-achievement-gains-3/

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research, 51*(6), 881–898. doi:10.1080/00273171.2016.1228042

Kline, R. B. (2015). Principles and practice of structural equation modeling. New York, NY: Guilford..

Kopriva, R. J., Thurlow, M. L., Perie, M., Lazarus, S. S., & Clark, A. (2016). Test takers and the validity of score interpretations. *Educational Psychologist, 51*(1), 108-128. doi:10.1080/00461520.2016.1158111

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234–249. doi:10.3102/0013189x17718797

Kuhfeld, M. (2016). *Multilevel item factor analysis and student perceptions of teacher effectiveness* (Doctorial dissertation). Retrieved from ProQuest Dissertations & Theses Global. (ED589875)

Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the Tripod student survey. *Educational Assessment, 22*(4), 253–274. doi:10.1080/10627197.2017.1381555

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education, 36*, 143–152. doi:10.1016/j.tate.2013.07.010

Kyriakides, L., & Creemers, B. (2008). Using a multidimensional approach to measure the impact of classroom-level factors upon student achievement: A study testing the validity of the dynamic model. *School Effectiveness and School Improvement, 19*(2), 183–205. doi:10.1080/09243450802047873

Kyriakides, L., & Creemers, B. (2009). The effects of teacher factors on different outcomes: two studies testing the validity of the dynamic model. *Effective Education, 1*(1), 61–85. doi:10.1080/19415530903043680

Kyriakides, L., Creemers, B., Antoniou, P., & Demetriou, D. (2010). A synthesis of studies searching for school factors: Implications for theory and research. *British Educational Research Journal, 36*(5), 807–830. doi:10.1080/01411920903165603

Lacireno-Paquet, N., Morgan, C., & Mello, D. (2014). How states use student learning objectives in teacher evaluation systems: A review of state websites. REL 2014–013. *Regional Educational Laboratory Northeast & Islands*. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2014013.pdf.

Lam, A. C., Schenke, K., Conley, A. M., Ruzek, E. A., & Karabenick, S. A. (2015).

Student perceptions of classroom achievement goal structure: Is it appropriate to aggregate? *Journal of Educational Psychology, 107*(4), 1102-–1115. doi:10.1037/edu0000028

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936–949. doi:10.3758/s13428-015-0619-7

Liang, X., & Yang, Y. (2014). An evaluation of WLSMV and Bayesian methods for confirmatory factor analysis with categorical indicators. *International Journal of Quantitative Research in Education, 2*(1), 17–38. doi:10.1504/ijqre.2014.060972

Liu, K., Lindsay, J., Springer, J., Wan, Y., & Stuit, D. *The utility of teacher and student surveys in principal evaluations: An empirical investigation.* Retrieved from https://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/REL_2015047.pdf

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120-131. doi:10.1016/j.cedpsych.2008.12.001

Mabin Jr., T. B. (2016). *Student-teacher connection, race, and relationships to academic achievement* (Doctorial dissertation). Retrieved from ProQuest Dissertations & Theses Global.

MacCallum, R.C.. (2003.) 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *8*(1), 113-139. doi:10.1207/s15327906mbr3801_5

Marsh, H. W., Dicke, T., & Pfeiffer, M. (2019). A tale of two quests: The (almost) non-overlapping research literatures on students' evaluations of secondary-school and university teachers. *Contemporary Educational Psychology, 58*, 1–18. doi:10.1016/j.cedpsych.2019.01.011

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85–110. doi:10.1146/annurev-clinpsy-032813-153700

Martínez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance. *Educational Evaluation and Policy Analysis, 38*(4), 738–756. doi:10.3102/0162373716666166

McDonnell, L. M. (2015). Stability and change in Title I testing policy. *Russell Sage Foundation, 1*(3), 170–186. doi:10.7758/rsf.2015.1.3.09

McGuinn, P. (2012). Stimulating reform: Race to the Top, competitive grants and the Obama education agenda. *Educational Policy, 26*(1), 136–159. doi:10.1515/for-2014-0017

Measures of Effective Teaching longitudinal database. (2012). Retrieved from http://www.icpsr.umich.edu/icpsrweb/METLDB/index.jsp.

Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *Journal of Experimental Education, 82*(2), 143–167. doi:10.1080/00220973.2013.769412

Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling.

In J.Osborne (Ed.), *Best practices in quantitative methods* (pp. 488–508). Thousand Oaks, CA: Sage

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art–teacher effectiveness and professional learning. *School Effectiveness and School Improvement, 25*(2), 231–256. doi:10.1080/09243453.2014.885451

Murphy, J., Hallinger, P., & Heck, R. H. (2013). Leading via teacher evaluation the case of the missing clothes. *Educational Researcher, 42*(6), 349–354. doi:10.3102/0013189x13499625

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*(4), 338–354. doi:10.1111/j.1745-3984.1991.tb00363.x

Nachtigall, C., Kroehne, U., Funke, F., & Steyer, R. (2003). (Why) Should we use SEM?-Pros and cons of structural equation modelling. *Methods of Psychological Research Online, 8*(2), 1-22. Retrieved from https://www.semanticscholar.org/paper/(Why)-Should-We-Use-SEM-Pros-and-Cons-of-Structural-Nachtigall-Kr%C3%B6hne/db5942456d12909877d62cc971395bab251d31e0

Nilsen, T., & Gustafsson, J.-E. (2016). *Teacher quality, instructional quality and student outcomes*: New York, NY: Springer.

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review, 82*(1), 123–141. doi:10.17763/haer.82.1.v40p0833345w6384

Pennington, K. & Mead, S. (2016, December). *For good measure?: Teacher evaluation policy in the ESSA era.* Washington, DC: Bellwether Education Partner. Retrieved from http://bellwethereducation.org/publication/good-measure-teacher-evaluation-policy-essa-era.

Ravitch, D. (2013). Holding education hostage. *Educational Studies*(2), 281–284. Retrieved from https://ideas.repec.org/a/nos/voprob/2013i2p281-284.html

Reddy, L. A., Dudek, C. M., Peters, S., Alperin, A., Kettler, R. J., & Kurz, A. (2018). Teachers' and school administrators' attitudes and beliefs of teacher evaluation: A preliminary investigation of high poverty school districts. *Educational Assessment, Evaluation and Accountability, 30*(1), 47–70. doi:10.1007/s11092-017-9263-3

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. *CBE-Life Sciences Education, 15*(1), 1–9. doi:10.1187/cbe.15-08-0183

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement, 25*(2), 197–230. doi:10.1080/09243453.2014.885450

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. doi:10.1037/a0029315

Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *The American Economic Review, 100(*2), 261–266. doi:10.1257/aer.100.2.261

Ross, E., & Walsh, K. (2019). *State of the States 2019: Teacher and principal evaluation policy*. Retrieved from https://www.nctq.org/publications/State-of-the-States-2019:-Teacher-and-Principal-Evaluation-Policy

Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology, 5*. doi:10.3389/fpsyg.2014.00081

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling, 16*(4), 583–601. doi:10.1080/10705510903203466

Ruelens, A., Meuleman, B., & Nicaise, I. (2018). Examining measurement isomorphism of multilevel constructs: The case of political trust. *Social Indicators Research, 140*(3), 907–927. doi:10.1007/s11205-017-1799-6

Scheerens, J. (2015). Theories on educational effectiveness and ineffectiveness. *School Effectiveness and School Improvement, 26*(1), 10–-31. doi:10.1080/09243453.2013.858754

Scherer, R., & Gustafsson, J.-E. (2015). Student assessment of teaching as a source of information about aspects of teaching quality in multiple subject domains: an application of multilevel bifactor structural equation modeling. *Frontiers in Psychology, 6,* 1–15. doi:10.3389/fpsyg.2015.01550

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology, 7*. doi:10.3389/fpsyg.2016.00110

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM, 48*(1), 29–40. doi:10.1007/s11858-016-0765-0

Schneider, B., Grogan, E., & Maier, A. (2011). Improving teacher quality: A sociological presage. In M. T. Hallinan (Ed.), *Frontiers in sociology of eEducation* (pp. 163–180). Dordrecht: Springer Netherlands.

Schulz, J., Sud, G., & Crowe, B. (2014). Lessons from the field: The role of student surveys in teacher evaluation and development. Bellweather Education Partners. Retrieved from https://bellwethereducation.org/sites/default/files/Bellwether_StudentSurvey.pdf

Schweig, J. (2014a). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis, 36*(3), 259–280. doi:10.3102/0162373713509880

Schweig, J. (2014b). Quantifying error in survey measures of school and classroom

environments. *Applied Measurement in Education, 27*(2), 133-157. doi:10.1080/08957347.2014.880442

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. doi:10.3102/0034654307310317

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–23. doi:10.17763/haer.57.1.j463w79r56455411

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), 598–642. doi:10.3102/0034654313496870

Stapleton, L. M., McNeish, D. M., & Yang, J. S. (2016). Multilevel and single-level models for measured and latent variables when data are clustered. *Educational Psychologist, 51*(3/4), 317–330. doi:10.1080/00461520.2016.1207178

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis, 38*(2), 293-317. doi:10.3102/0162373715616249

Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics* (5th ed). Boston, MA: Pearson.

Tay, L., Woo, S. E., & Vermunt, J. K. (2014). A conceptual and methodological framework for psychometric isomorphism: Validation of multilevel construct

measures. *Organizational Research Methods, 17*(1), 77–106. doi:10.1177/1094428113517008

Taylor, J., Stecher, B., O'Day, J., Naftel, S., & Le Floch, K. C. (2010). *State and local implementation of the No Child Left Behind Act. Vol. IX--Accountability under NCLB: Final Report*. Washington, DC: U.S Department of Education. Retrieved from: https://files.eric.ed.gov/fulltext/ED508912.pdf

Thomas, J. Y., & Brady, K. P. (2005). Chapter 3: The Elementary and Secondary Education Act at 40: Equity, accountability, and the evolving federal role in public education. *Review of Research in Education, 29*(1), 51–67. doi:10.3102/0091732x029001051

Tripod frequently asked questions. (2017). Tripod Partners. Retrieved from Tripod website: http://tripoded.com/districts-states

U.S. Department of Education. (2009). *Race to the top program: Executive summary.* Washington, DC: Author. Retrieved from https://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Van der Schaaf, M., Slof, B., Boven, L., & De Jong, A. (2019). Evidence for measuring teachers' core practices. *European Journal of Teacher Education, 42*(5), 675-694. doi:10.1080/02619768.2019.1652903

van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement, 30*(1), 30–50. doi:10.1080/09243453.2018.1539015

Van de Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology, 6,* 1064. doi: 10.3389/fpsyg.2015.01064.

van der Steeg, M., & Gerritsen, S. (2016). Teacher evaluations and pupil achievement gains: Evidence from classroom observations. *De Economist*, 1–25. doi:10.1007/s10645-016-9280-5

Van Mierlo, H., Vermunt, J. K., & Rutte, C. G. (2009). Composing group-level constructs from individual-level survey data. *Organizational Research Methods, 12*(2), 368–392. doi:10.1177/1094428107309322

Vanlaar, G., Kyriakides, L., Panayiotou, A., Vandecandelaere, M., McMahon, L., De Fraine, B., & Van Damme, J. (2016). Do the teacher and school factors of the dynamic model affect high-and low-achieving student groups to the same extent? a cross-country study. *Research Papers in Education, 31*(2), 183–211.

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod Student Perception Survey. *American Educational Research Journal, 53*(6), 1834–1868. doi:10.1080/02671522.2015.1027724

White, M., & Rowan, B. (2014). *User guide to the Measures of Effective Teaching Longitudinal Database (MET LDB).* Ann Arbor: Inter-University Consortium for Political and Social Research, University of Michigan. Retrieved from https://www.icpsr.umich.edu/icpsrweb/METLDB/

Wang, X., & Lee, S. (2018). Validating a new survey instrument measuring factors contributing to transfer in STEM. *The Review of Higher Education*, *42*(2), 339–384. doi:10.1353/rhe.2019.0000

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. Brooklyn, NY: New Teacher Project. Retrieved from https://tntp.org/assets/documents/TheWidgetEffect_execsummary_2nd_ed.pdf

Wirth, R., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods, 12*(1), 58–79. doi:10.1037/1082-989x.12.1.58

Woods, J. (2015). *Instructional time trends. Education trends. Education Commission of the States.* Retrieved from: https://files.eric.ed.gov/fulltext/ED558372.pdf

Zhang, Z., Zhang, Z., Lee, J. C.-K., Lee, J. C.-K., Wong, P. H., & Wong, P. H. (2016). Multilevel structural equation modeling analysis of the servant leadership construct and its relation to job satisfaction. *Leadership & Organization Development Journal, 37*(8), 1147–1167. doi:10.1108/lodj-07-2015-0159

Zumbo, B. D., & Forer, B. (2011). Testing and measurement from a multilevel view: Psychometrics and validation. In J. A. Bovaird, K. F. Geisinger, C. W. Buckendahl, J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K–12 settings* (pp. 177–190). Washington, DC: American Psychological Association.

Appendix A: MPlus Code for Model Estimation.

Model 1

```
TITLE:      Model 1, Single Factor Measurement Model
VARIABLE:   Names are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is complex;
            Estimator is WLSMV
MODEL:      Tripod by c1-c36;
OUTPUT:     Stdyx;
```

Note: see Figure 2.

Model 2

```
TITLE:      Model 2, Three Factor Measurement Model
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Cluster is section;
ANALYSIS:   Type is complex;
            Estimator is WLSMV;
MODEL:      Psupport by c1-c3 r_c27-c31;
            Asupport by c4-c10 c16-c22;
            Csupport by c11-r_c15 c23-c26;
OUTPUT:     Stdyx;
```

Model 3

```
TITLE:      Model 3, Seven-Factor Factor Measurement Model
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is complex;
            Estimator is WLSMV;
MODEL:      Care by c1-c3;
            Control by c4-r_c10;
            Clarify by c11-r_c15;
            Challenge by c16-c22;
            Captivate by c23-c31;
            Confer by r_c27-c31;
            Consolidate by c32-c36;
OUTPUT:     Stdyx;
```

Note: see Figure 4

Model 4

```
TITLE:      Model 4, 7-Factor ESEM
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is complex;
            Estimator is WLSMV;
            Rotation is Target;
MODEL:      Care by c1-c3 c4-c36~0 (*1);
            Control by c4-R_c10 c1-c3~0 c11-c36~0 (*1);
            Clarify by c11-r_c15 c1-r_c10~0 c16-c36~0 (*1);
            Challenge by c16-c23 c1-r_c15~0 c24-c36~0 (*1);
            Captivate by c24-r_c27 c1-c23~0 c28-c36~0(*1);
            Confer by c28-c32 c1-r_c27~0 c33-c36~0 (*1);
            Consolidate by c33-c36 c1-c32~0 (*1);
OUTPUT:     Stdyx;
```

Model 5

```
TITLE:      Reduced Form 2-Factor Tripod
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is complex;
            Estimator is WLSMV;
MODEL:      Clar by C11-R_C15; !(1-5);
            Cap by  C24-R_C27; !(6-9);
            Resid1 by R_C15@1;
            Resid2 by R_C27@1;
OUTPUT:     Stdyx;
```

Note: see Figure 11

Model 6

```
TITLE:      One 2nd Order Factor
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is complex;
            Estimator is WLSMV;
MODEL:      Care by c1 c2 c3;
            Control by c4 c5-r_c10;
            Clarify by c11 c12-r_c15;
            Challenge by c16 c17-c22;
            Captivate by c23 c24-c26;
            Confer by r_c27 c28-c31;
            Consolidate by c32 c33-c36;
            General by Care Control Clarify Challenge Captivate
            Confer Consolidate;
OUTPUT:     Stdyx;
```

Model 7

```
TITLE:      Three 2nd Order Factor Model
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is complex;
            Estimator is WLSMV;
MODEL:      Care by c1 c2 c3;
            Control by c4 c5-r_c10;
            Clarify by c11 c12-r_c15;
            Challenge by c16 c17-c22;
            Captivate by c23 c24-c26;
            Confer by r_c27 c28-c31;
            Consolidate by c32 c33-c36;
            PSupport by Care Confer;
            ASupport by Control Challenge;
            Csupport by Captivate Clarify Consolidate;
OUTPUT:     Stdyx;
```

Note: see Figure 7.

Model 8

```
TITLE:      One 3rd Order, Three 2nd Order Factors Model
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is complex;
            Estimator is WLSMV;
MODEL:      Care by c1 c2 c3;
            Control by c4 c5-r_c10;
            Clarify by c11 c12-r_c15;
            Challenge by c16 c17-c22;
            Captivate by c23 c24-c26;
            Confer by r_c27 c28-c31;
            Consolidate by c32 c33-c36;
            PSupport by Care Confer;
            ASupport by Control Challenge;
            Csupport by Captivate Clarify Consolidate;
            General by PSupport;
            General by ASupport;
            General by CSupport;

OUTPUT:     Stdyx;
```

Note: see Figure 6.

Model 9

```
TITLE:      Tripod Bifactor Model
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is complex;
            Estimator is WLSMV;
MODEL:      Care by c1 c2 c3;
            Control by c4 c5-r_c10;
            Clarify by c11 c12-r_c15;
            Challenge by c16 c17-c22;
            Captivate by c23 c24-c26;
            Confer by r_c27 c28-c31;
            Consolidate by c32 c33-c36;
            Gen by c1-c36;
            Gen with Care-Consolidate @0;
            Care with Control-Consolidate@0;
            Control with Clarify-Consolidate@0;
            Clarify with Challenge-Consolidate@0;
            Challenge with Captivate-Consolidate@0;
            Captivate with Confer-Consolidate@0;
            Confer with Consolidate@0;
OUTPUT:     Stdyx;
```

Model 10

```
TITLE:      Three Factor Bifactor Model
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is complex;
            Estimator is WLSMV;
MODEL:      Gen by c1-c36;
            PSupport by c1 c2 c3 r_c27-c31;
            CSupport by c11 c12-r_c15 c23-c26 c32-c36;
            ASupport by c4 c5-r_c10 c16-c22;
            Gen with PSupport CSupport ASupport @0;
            PSupport with CSupport ASupport @0;
            CSupport with ASupport @0;
OUTPUT:     Stdyx;
```

Model 11

```
TITLE:      Wallace Bifactor Model
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is complex;
            Estimator is WLSMV;
MODEL:      Gen by c1-c36;
            Control with c4-r_c10;
            Reverse with r_c8-r_c10 r_c15 r_c27;
            Gen with Control-Reverse@0;
            Control with Reverse@0;
OUTPUT:     Stdyx;
```

Model 12

```
TITLE:      Wallace Hybrid Bifactor Model
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is complex;
            Estimator is WLSMV;
MODEL:      Gen by c1-c36;
            Clarify by c11-r_c15;
            Captivate by c24-r_27;
            Gen with Clarify-Captivate@0;
            Clarify with Captivate@0;
OUTPUT:     Stdyx;
```

Multilevel Models

Model 3

```
TITLE:      ML Tripod
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;

ANALYSIS:   Type is twolevel;
            Estimator is WLSMV;
MODEL:      %within%
            Carew by c1 c2 c3;
            Controlw by c4-R_c10;
            Clarifyw by c11-r_c15;
            Challengew by c16-c22;
            Captivatew by c23-c26;
            Conferw by r_c27-c31;
            Consolidatew by c32-c36;
            %between%
            Careb by c1 c2 c3;
            Controlb by c4-R_c10;
            Clarifyb by c11-r_c15;
            Challengeb by c16-c22;
            Captivateb by c23-c26;
            Conferb by r_c27-c31;
            Consolidateb by c32-c36;
OUTPUT:     Stdyx;
```

Model 5

```
TITLE:      ML 2 Factor Tripod with Residual Factor
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;

ANALYSIS:   Type is twolevel;
            Estimator is WLSMV;
MODEL:      %within%
            Clarw by C11-R_C15;
            Capw by  C24-R_C27;
            Residw by R_C15@1;
            Residw by R_C27@1;
            Residw with Clarw@0;
            Residw with Capw@0;
            %between%
            Clarb by C11-R_C15;
            Capb by C24-R_C27;
            Residb by R_C15@1;
            Residb by R_C27@1;
            Residb with Clarb@0;
            Residb with Capb@0;
OUTPUT:     Stdyx;
```

Model 11

```
TITLE:      ML Wallace Bifactor Model
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is twolevel;
            Estimator is WLSMV;
MODEL:      %within%
            Genw by c1-c36
            ControlW by c4-r_c10;
            ReverseW by r_c8-r_c10 r_c15 r_c27;
            GenW @1;
            ControlW@1;
            ReverseW@1;
            GenW with ControlW@0;
            GenW with ReverseW@0;
            ControlW with ReverseW@0;
            %between%
            GenB by c1-c36
            ControlB by c4-r_c10;
            ReverseB by r_c8-r_c10 r_c15 r_c27;
            GenB @1;
            ControlB@1;
            ReverseB@1;
            GenB with ControlW@0;
            GenB with ReverseW@0;
            ControlB with ReverseW@0;
OUTPUT:     Stdyx;
```

Model 12

```
TITLE:      ML Wallace Hybrid Bifactor Model
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;
ANALYSIS:   Type is twolevel;
            Estimator is WLSMV;
MODEL:      %within%
            Genw by c1-c36;
            Captivatew by c24-r_c27;
            Clarifyw by c11-r_c15
            Controlw by c4-r_c10;
            Reversew by r_c8 r_c9 r_c10 r_c15 r_c27;
            Genw with Controlw@0;
            Genw with Reversew@0;
            Genw with Captivatew@0;
            Genw with Clarifyw@0;
            Controlw with Reversew@0;
            Controlw with Captivatew@0;
            Controlw with Clarifyw@0;
            Reversew with Captivatew@0;
            Reversew with Clarifyw@0;
            Captivatew with Clarifyw@0;
            %between%
            Genb by c1-c36;
            Clarfiyb by c11-r_c15;
            Captivateb by c24-r_c27;
            Controlb by c4-r_c10;
            Reverseb by r_c8 r_c9 r_c10 r_c15 r_c27;
            Genb with Controlb@0;
            Genb with Reverseb@0;
            Genb with Captivateb@0;
            Genb with Clarifyb@0;
            Controlb with Reverseb@0;
            Controlb with Captivateb@0;
            Controlb with Clarifyb@0;
            Reverseb with Captivateb@0;
            Reverseb with Clarifyb@0;
            Captivateb with Clarifyb@0;
OUTPUT:     Stdyx;
```

Metric Isomorphism

Model 5

```
TITLE:      ML Model 5 Equal Factor Loadings
VARIABLE:   Name are y1-y36;
            Categorical are y1-y36;
            Missing are all (-9);
            Cluster is section;

ANALYSIS:   Type is twolevel;
            Estimator is WLSMV;
MODEL:      %within%
            Clarw by c11* c12-r_c15 (1-5);
            Capw by c24* c25-r_c27 (6-9);
            Clarw@1;
            Capw@1;
            Residw by r_c15@1;
            Residw by r_c27@1;
            Residw with Clarw@0;
            Residw with Capw@0;
            %between%
            Clarb by c11* c12-r_c15 (1-5);
            Capb by c24* c25-r_c27 (6-9);
            Clarb;
            Capb;
            c11-r_c15@0;
            c24-r_c27@0;
            Residb by r_c15@1;
            Residb by r_c27@1;
            Residb with Clarb@0;
            Residb with Capb@0;
OUTPUT:     Stdyx;
```

Appendix B: Years 1 and 2 Level-Specific Fit Calculations

			Year 1		Year 2	
Model	Between	Within	Chi-Square	DF	Chi-Square	DF
Model 3a	H0	SAT	13506.088	573	5073.289	573
Model 3b	SAT	HO	35848.444	573	35848.444	573
Model 3c	IND	SAT	63177.749	630	63177.749	630
Model 3d	SAT	IND	1188003.479	630	1188003.479	630
Model 5a	H0	SAT	93.689	26	115.756	26
Model 5b	SAT	HO	2860.390	26	1288.154	26
Model 5c	IND	SAT	12852.275	36	4837.410	36
Model 5d	SAT	IND	386300	36	156481.466	36
Model 5RFa	H0	SAT	98.936	25	118.961	25
Model 5RFb	SAT	HO	1226.560	25	611.124	25
Model 5RFc	IND	SAT	12852.275	36	4837.410	36
Model 5RFd	SAT	IND	386300.969	36	156481.466	36
Model 5RFa*	H0	SAT	108.10	25	118.961	25
Model 5RFb*	SAT	HO	976.719	25	219.760	25
Model 5RFc*	IND	SAT	5189.112	36	2883.491	36
Model 5RFd*	SAT	IND	45230.573	36	151531.96	36
Model 11a	H0	SAT	14244.829	582	5984.333	582
Model 11b	SAT	HO	44467.201	582	18161.755	582
Model 11c	IND	SAT	159074.04	630	63177.749	630
Model 11d	SAT	IND	3001063.81	630	1188003.479	630
Model 12a	H0	SAT	13767.660	573	5722.238	573
Model 12b	SAT	H0	33441.561	573	13428.525	573
Model 12c	IND	SAT	159074.04	630	63177.749	630
Model 12d	SAT	IND	3001063.81	630	1188003.479	630
Model 5 RF	By Subject					
Biology	H0	SAT			37.915	25
	SAT	HO			278.682	25
	IND	SAT			1767.492	36
	SAT	IND			51156.234	36
English	H0	SAT			28.897	25
	SAT	HO			248.581	25
	IND	SAT			1579.930	36
	SAT	IND			58372.021	36
Mathematics	H0	SAT			81.378	25
	SAT	HO			168.648	25
	IND	SAT			293.438	36
	SAT	IND			21379.271	36

Estimator = WLSM; RF = residual factor; H0 = hypothesized model; SAT = saturated model; IND = independence model

Appendix C: Year 2 Test of Model Fit, Single Level

Models		χ^2	*df*	RMSEA (90% CI)	CFit	CFI	SRMR
Individual Factors							
Care	C1-C3	26,610.210*	0	-	-	-	-
Control	C4-C10	1,864.201	14	.128 (.123-.133)	0.0	.898	.034
Clarify	C11-C15	1,864.201	5	.060 (.052-.069)	.018	.996	.010
Challenge	C16-C23	1,221.329	20	.086 (.082-.090)	0.0	.974	.024
Captivate	C24-C27	75.348	2	.067 (.055-.081	.012	.997	.006
Confer	C28-C32	177.173	5	.065 (.057-.074)	.001	.984	.016
Consolidate	C33-C36	281.985	2	.132 (.119-.145)	0.0	.979	.018
Hypothesized Factor Structures							
Model 0	Baseline	191,604.758	630	.268			
Model 1	1 Factor	34,862.400	594	.084(.084-.085)	0.0	.719	.700
Model 2:	3 Factor	31,461.814	591	.080 (.079-.081)	0.0	.747	.067
Model 3	7 Factor	15,029.210	573	.056 (.055-.057)	0.0	.882	.044
Model 4	7-Factor ESEM	3,438.899	399	.031 (.030-.032)	1.00	.975	.012
Model 5	2-Factor	598.323	25	.053 (.049-.057)	.078	.987	.013

Note. N = 8,120; with only three indicators the model cannot be tested; χ^2 = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; CFit = close fit (probability that RMSEA < .05); CFI = comparative fit index. SRMR = standardized root mean square residual.

Appendix D: Year 2 Test of Model Fi, Higher Order Factor Structure

Models		χ^2	df	RMSEA (90% CI)	CFit	SRMR	CFI
Hypothesized Factor Structure							
Model 6	1-2nd Order Factor	14,679.565	587	.054 (.054-.055)	0.00	.045	.884
Model 7	3-2nd Order Factors	14,977.633	584	.055 (.054-.056)	0.00	.045	.882
Model 8	1-3rd 3-2nd Order	14,977.629	584	.055 (.054-.056)	0.00	.045	.822
Bifactor Models							
Model 9	7C Bifactor	Did not converge					
Model 10	3C Bifactor	18,897.137	584	.064 (.063-.064)	0.00	.045	.850
Model 11	Wallace Bifactor	10,058.477	582	.045 (.044-.046)	.1.00	.032	.922
Model 12	Wallace2 Bifactor	7,848.117	573	.040 (.039-.404)	1.00	.029	.940

Note. N = 8,120; with only three indicators the model cannot be tested; χ^2 = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; CFit = close fit (probability that RMSEA < .05); CFI = comparative fit index. SRMR = standardized root mean square residual.