Title
Multilevel Item Factor Analysis and Student Perceptions of Teacher Effectiveness

Author
Kuhfeld, Megan Rebecca

Publication Date
2016

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Multilevel Item Factor Analysis and Student Perceptions of Teacher Effectiveness

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Education

by

Megan Rebecca Kuhfeld

2016

Abstract of the Dissertation

Multilevel Item Factor Analysis and Student Perceptions of Teacher Effectiveness

by

Megan Rebecca Kuhfeld

Doctor of Philosophy in Education

University of California, Los Angeles, 2016

Professor Li Cai, Chair

Measures of teacher effectiveness have become a major research and policy issue due to the increased focus on teacher accountability during the past decade. Growing concerns about the variability in the quality of teaching and traditional approaches to measuring teacher effectiveness led to federal and state policies calling for more rigorous measures of teacher effectiveness (Kane & Cantrell, 2010; Weisberg et al., 2009). One of the increasingly used teacher effectiveness measures is student surveys of instructional practice. These surveys are now being given in grades K-12 for accountability purposes, to provide teachers with feedback to improve their teaching, and to guide professional development (Bill & Melinda Gates Foundation, 2012). Given student surveys are widely used to assess and improve teacher effectiveness, it is important to examine the reliability and validity of these measures.

This dissertation focused on the secondary Tripod Survey, which is the most widely used off-the-shelf student survey instrument for use in middle and high schools (Ferguson, 2010). The Tripod survey asks students to provide feedback on teacher practices and student behavior, which are operationalized as the Tripod 7Cs framework of teacher effectiveness. The seven domains are Care, Control, Clarify, Challenge, Captivate, Confer and Consolidate (Ferguson, 2012). According to the survey developer, over 100,000 teachers have received feedback using Tripod surveys (Tripod Project, 2016). Despite this widespread use, little has been published regarding the psychometric properties of the instrument (Camburn, 2012).

In this dissertation, I describe an innovative methodological approach for exploring the dimensionality and collecting validity evidence to support the use of the Tripod survey as a measure of teacher quality. This approach uses a multilevel extension of full-information item factor analysis models. Item factor analysis (IFA) models are widely used in educational measurement research (Wirth & Edwards, 2007), though these models have traditionally ignored the hierarchical, nested structure of educational systems and treated all individuals as independent. Multilevel IFA models enable the data to be treated in an appropriate manner, with item responses nested within individuals within groups. However, these models are computationally intensive, and until recently were not available in commercial software. With the development of the Metropolis–Hastings Robbins–Monro (MH-RM) algorithm (Cai, 2010b), which allows for the estimation of high-dimensional IFA models, and the implementation of multilevel IFA models in the item response modeling software flexMIRT® (Cai, 2015), these models can now be readily applied to educational research questions.

The aims of this dissertation are two-fold. First, I provide an introduction to the multilevel item factor analysis (IFA) modeling framework, and demonstrate the flexibility and efficiency of this model in various educational settings. It is essential to establish that the multilevel IFA model can be estimated under realistic data conditions prior to using this modeling technique to answer important educational policy questions regarding student surveys. Secondly, I use multilevel IFA models to examine the dimensionality, reliability, and validity of the Tripod secondary student survey.

More specifically, I investigate the following research questions:

1. Can I efficiently and accurately estimate multilevel IFA models in the context of educational assessment and survey data?

2. Is possible to detect sources of model misfit in multilevel IFA models using a newly developed goodness-of-fit statistic?

3. Can I use the multilevel IFA model to produce estimates of teacher practice scores that clarify the degree to which the seven dimensions of teacher practice measured by the Tripod survey simultaneously predict student learning?

4. Using data from six urban school districts collected by the Measures of Effective Teaching (MET) Project, is there validity evidence that supports the use of the Tripod survey for summative and formative teacher evaluation purposes?

The findings from this dissertation contribute to methodological and substantive bodies of work. Methodologically, I demonstrate that the multilevel IFA model can be used to make reliable group-level inferences across a variety of educational contexts. Additionally, I propose a limited-information goodness-of-fit statistic for multilevel IFA models to address the current limitation of these models that there is no established consensus on how to assess the model fit.

In addition, this dissertation contributes to the field of teacher evaluation by analyzing the validity of the secondary Tripod survey. This work represents the first systematic review of the psychometric and validity properties of the Tripod survey. The findings call into question whether the current practice of reporting feedback in terms of the 7Cs is warranted. In particular, the gathered evidence does not support distinguishing among the six of the 7Cs teacher practices (Care, Clarify, Consolidate, Confer, Challenge, and Captivate). Therefore, I propose combining the items from these sub-domains into a single Teacher Support scale. Both Support and Control scores are found to be related to teacher observation scores, but only teachers' level of Control is predictive of student achievement. In summary, this study provides promising evidence that the widely used Tripod survey is a useful tool for measuring two important dimensions of teacher effectiveness.

The dissertation of Megan Rebecca Kuhfeld is approved.

Meredith Phillips

Mark Hansen

Mike Seltzer

Li Cai, Committee Chair

University of California, Los Angeles

2016

For my parents.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This work could not have been completed without the support and guidance of my advisor and committee chair, Professor Li Cai. I have learned so much from him, and feel very fortunate to have benefited from his guidance and wisdom during my time at UCLA. In addition, the dissertation made extensive use of flexMIRT® item response modeling software, which Professor Cai developed. I am also thankful to the other members of my committee, Dr. Meredith Phillips, Dr. Mike Seltzer, and Dr. Mark Hansen, for their mentorship throughout my entire graduate career. I am fortunate to have worked with Meredith Phillips and Kyo Yamashiro on the Los Angeles Classroom and School Environment Survey during my first years at UCLA, which prompted a great deal of my interest in measures of teacher effectiveness. Furthermore, I would like to thank Dr. Scott Monroe for his assistance with technical components of this dissertation, and for sharing his work on cluster student growth percentiles.

I am very grateful for the support and feedback from past and current colleagues at UCLA: Carl Falk, Jon Schweig, Alex Sturm, Talia Stol, Lisa Dillman, Alejandra Priede, Kat Schenke, and Larry Thomas. I am also thankful to my SRM cohort: Jane Li, Kevin Schaaf, Jenn Ho, Jason Tsui, Danny Dockterman, & Liz Perez. I am very fortunate to have shared my Ph.D. experience with you all.

I wish to acknowledge financial support for this research and my training that has been provided by the Institute of Education Sciences (through a predoctoral training grant awarded to UCLA, R305B080016), the AERA-MET Dissertation Fellowship, and the UCLA Department of Education. The views expressed here are my own and do not reflect the views or policies of these funding agencies.

Finally, I want to thank my parents for their unwavering support and inspiration throughout all of my years of school. Lastly, I thank Philip for his support and for helping me to stay grounded throughout the entire Ph.D. process. Without his support this would not have been possible.

Vita

Education

2009	Bachelors of Science, Duke University, Psychology.
2012	Master's of Arts, University of California, Los Angeles. Education.
2016	Master's of Sciences, University of California, Los Angeles. Statistics.

Work

2009–11	Senior Research Assistant in Education, Child Trends.
2011–2013	Graduate Student Researcher, Los Angeles Education Research Institute.
2011–present	Graduate Student Researcher, University of California, Los Angeles.

Publications

Stucky, B.D., Edelen M.O., Tucker, M.S, Shadel, W.G., Cerully, J., Kuhfeld, M., Hansen, M., & Cai, L. (2014). Development of the PROMIS® negative psychosocial expectancies item bank. *Nicotine & Tobacco Research, 16*(3), S241-S249.

Tucker, M.S, Shadel, W.G., Edelen M.O., Stucky, B.D., Kuhfeld, M., Hansen, M., & Cai, L. (2014). Development of the PROMIS® social motivations item bank. *Nicotine & Tobacco Research, 16*(3), S241-S249.

Lippman, L., Moore, K. A., Guzman, L, Ryberg, R., McIntosh, H., Caal, S., Ramos, M., Carle, A., & Kuhfeld, M. (2014). *Flourishing children: Defining and testing indicators of positive development.* Heidelberg, Germany: Springer.

Kuhfeld, M. (2015). An Interpretive Validity Analysis of the Tripod Survey. *Multivariate Behavioral Research*, 50(6), (Abstract).

Lee, T., Cai, L., & Kuhfeld, M. (2015). A poor person's posterior predictive checking of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(2), 206–220.

Cai, L., Choi, K., & Kuhfeld, M. (2016). On the role of multilevel item response models in multi-site evaluation studies for serious games. In O'Neil, H. F., Baker, E. L., & Perez, R. (Eds.) *Using Games and Simulations for Teaching and Assessment.* New York, NY: Taylor & Francis.

CHAPTER 1

Introduction

Measures of teacher effectiveness have been a central focus of educational policy in the last decade. One of the central reasons is the recent body of research documenting classroom-to-classroom variation in students' academic achievement (see, for example, Nye, Konstantopoulos, & Hedges, 2004; Rockoff, 2004; Kane & Staiger, 2008). It has been estimated that out-of-school factors explain 60 percent of the variance in student test scores (Goldhaber, Brewer, & Anderson, 1999). However, educators and policy-makers often can have very little influence on out-of-school factors, so researchers have focused on identifying the within-school influences that have the most impact on student learning. Haertel's (2013) review of the research regarding sources of influence on student test scores concluded that approximately ten percent of the total variation in student test scores gains in a single year is due to variation among teachers. While this may seem small, it is the largest *within-school* influence on student learning.

Despite the recognition of the importance of effective teachers, it is widely acknowledged that school districts have traditionally been unsuccessful at recognizing highly effective teachers (Kane & Cantrell, 2010). Until recently, the factors that have been used for pay determination in school districts have mainly been "observable" teacher characteristics, such as years of experience, professional certification, and degree attainment (Hanushek & Rivkin, 2006). The 2001 No Child Left Behind Act (NCLB) defined "highly qualified" teachers of core subjects as those with at least a bachelor's degree, a state license, and demonstrated competency in the subject matter taught (Bush, 2001). However, some observable teacher qualifications, such as degree attainment and certification status, have not been found to be predictive of student outcomes (Rockoff, Jacob, Kane, & Staiger, 2011; Hanushek & Rivkin, 2006; Goldhaber & Hansen, 2013).

Furthermore, under traditional teacher evaluation systems, most teachers within a

school district were given the same rating. In a 2009 review of teacher evaluation systems in 12 districts and four states, Weisberg et al. (2009) showed that in districts that used binary evaluation ratings (generally "satisfactory" or "unsatisfactory"), more than 99 percent of teachers receive the satisfactory rating. Furthermore, districts with more than two categories did not show more equal distribution of ratings, with 94 percent of teachers receiving one of the top two ratings, and only one percent rated unsatisfactory. This is problematic for two reasons: (a) truly exceptional teachers cannot be recognized, compensated, or promoted when the majority of teachers are deemed good or great, and (b) ineffective teachers are not identified for professional development or additional supports. Toch and Rothman (2008) summarized the general opinion of traditional evaluation practices, stating that they are "superficial, capricious, and often don't even directly address the quality of instruction, much less measure students' learning" (p. 1).

Recent federal policies have catalyzed changes in teacher evaluation systems across the country. Race to the Top (RTTT), a competitive grant program that begun in 2009, called for states who were competing for million dollars worth of funding to establish more rigorous teacher evaluation systems that relied on multiple measures, with a strong emphasis on student growth (U.S. Department of Education, 2009). Additionally, in 2011, a program was started to provide waivers to the Elementary and Secondary Education Act (ESEA). In order to get a waiver, state education agencies were encouraged to develop teacher evaluation systems that emphasized the use of multiple measures, with a student growth used as a significant factor in the evaluation system (Popham, 2013). These federal policies were very successful at getting state education agencies to enact changes in teacher evaluation policies. According to the National Council of Teacher Quality, between 2009 and 2012, 36 states and the District of Columbia introduced new teacher evaluation policies (Doherty & Jacobs, 2013).

While there seems to be consensus that previous models of teacher evaluation were not functioning properly, it is difficult to find agreement on the definition of teaching quality, and how it should be measured. Race to the Top (RTTT) defines an effective teacher as "a teacher whose students achieve acceptable rates (e.g., at least one grade

level in an academic year) of student growth" (U.S. Department of Education, 2009, p. 12). At the other extreme, Bennett (2002, p. 2) states that "teaching is more than 'facilitating the acquisition of skills'. It is offering an invitation and encouragement to life, to a fulfilled life." These two definitions vary greatly in the scope of what is expected of teachers, and how the evaluation of teachers would be conceptualized.

It is also worth noting that there are often multiple intended goals of teacher evaluation systems (Popham, 2013). There is an important distinction between the intentions of *formative* and *summative* teacher evaluation. Popham (2013) defines *formative teacher evaluation* as evaluation activities that are intended toward improving teacher's instruction, while *summative teacher evaluation* is evaluation activities with the aim of making decisions about teachers that lead to future consequences (e.g., rewards or employment termination). Many teacher evaluation systems use scores from measures of teacher effectiveness for both summative and formative purposes, such as to provide feedback to teachers regarding strong and weak teaching characteristics, to identify struggling teachers for professional development, and to reward high performing teachers. Acknowledging the multiple purposes of teacher evaluation systems is important because a measure of teacher effectiveness may be found to be valid for one purpose but not another (Kane, 2006).

1.1 Multiple measures of teacher practice

An assortment of different measures have been proposed for assessing teacher quality and instructional practice. One central group of measures focuses on student academic growth, which can take many forms but are broadly grouped under the label "value-added models (VAM)". Examining student achievement gains after adjusting for some student and school characteristics leads to more fair comparisons of teachers than judgments based on their students' test scores at a single point in time. In a highly publicized report, teacher value-added scores predicted a classroom of students' long-term outcomes, including college attendance and future salaries (Chetty, Friedman, & Rockoff, 2011). However, the underlying assumptions and methodological flaws of value added models have been well-discussed and criticized in recent years (see,

for example Haertel, 2013; Baker et al., 2010; McCaffrey, 2012; American Statistical Association, 2014). The current consensus is that value-added model scores alone are "not sufficiently reliable and valid indicators of teacher effectiveness to be used in high-stakes personnel decisions" (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012, p. 2).

A second major group of teacher practice measures involves systematic classroom observation. Classroom observation is widely regarded as a key component of teacher evaluation systems (Popham, 2013). A variety of different observational measures have been developed for evaluating teachers, including the Classroom Assessment Scoring System (Pianta, La Paro, & Hamre, 2008) and the Framework for Teaching (Danielson, 2007). To be implemented reliably, a fair amount of training and monitoring for rater agreement is required (Joe, Tocci, Holtzman, & Williams, 2013), which is why classroom observations are often seen as prohibitively expensive (Rowan & Correnti, 2009).

Teacher logs, vignettes, and student work portfolios have also been proposed and tested mostly in research settings (Kennedy, 1999). These measures are generally time-consuming and expensive to collect. Teacher tests, such as the Mathematical Knowledge for Teaching Questionnaire (Hill, Schilling, & Ball, 2004), have been developed to measure teacher content knowledge, but cannot provide input on pedagogical skills or classroom interactions.

Student perception surveys, including the Tripod survey (Ferguson, 2010), measure student opinions about their teacher's practices and classroom environment. Student surveys are fairly cheap to administer, measure characteristics valued by teachers, and have been found to be predictive of student academic growth (Bill & Melinda Gates Foundation, 2012). These surveys are increasingly being incorporated into state evaluation systems (Doherty & Jacobs, 2015), and are the focus of this dissertation.

1.2 Student surveys of instructional practice

Student surveys of instructional practice have recently gained prominence among researchers and policymakers as an inexpensive way to get feedback on what is occurring

inside the classroom. The goal of these surveys is to provide fair and reliable feedback to teachers regarding their students' perceptions of the strengths and weaknesses of a range of teacher practices.

The Tripod survey (Ferguson, 2010) is the most widely-used off-the-shelf student survey instrument (Bill & Melinda Gates Foundation, 2012). The Tripod student perceptions assessment was developed by Ron Ferguson at Harvard University, and is based upon classroom-level surveys developed by the Tripod Project for School Improvement (Ferguson, 2010). The "tripod" describes the knowledge and skills that are needed to deliver instruction effectively: (a) content knowledge, (b) pedagogic knowledge and skills, and (c) the ability to connect with students on a personal level. The Tripod survey focuses primarily on what teachers do and how the classroom operates, which is operationalized as the Tripod 7Cs framework of teacher effectiveness. The seven domains are Care, Control, Clarify, Challenge, Captivate, Confer and Consolidate (Ferguson, 2012).

The Tripod survey is used in grades K-12 in school districts all over the country. The Tripod Project website states that "over 100,000 teachers have received valuable feedback from millions of students who have completed Tripod surveys" (Tripod Project, 2016). Schweig (2014) found that 75 percent of states and local districts that mention a specific survey instrument for use in their teacher evaluation systems are using the Tripod Survey. The Tripod was also used as a part of the Measures of Effective Teaching (MET) Project, a large study of teacher effectiveness funded by the Bill and Melinda Gates Foundation.

Student surveys are currently being used with multiple intents. The stated intention of the Tripod survey is to be both a diagnostic and professional development tool. Specifically, the survey can be used to provide information that will guide setting professional development and coaching goals (Tripod Project, 2016). In a brief about student surveys, MET project researchers state that "Teachers should receive their results in a timely manner, understand what they mean, and have access to professional development resources that will help them target improvement in areas of need" (Bill & Melinda Gates Foundation, 2012, p. 4). However, not a great deal is known about

the utility of the Tripod (or any other student survey) as a formative tool. Additionally, teacher-level scores from student surveys are a required component of summative evaluation scores in seven states, with 26 other states allowing for the use of student surveys in teacher evaluations (Doherty & Jacobs, 2015).

1.3 Advantages and disadvantages of student surveys

There are many advantages that have been cited for using student surveys as a part of a teacher evaluation system. Proponents of student surveys note that students are natural observers of the classroom in which they spend their days, and provide feedback that you cannot get a single classroom observation (Ferguson, 2012). The main advantages of student surveys cited by the MET Project are the fact that survey results (unlike value-added models) point to strengths and areas for improvement, the items have face validity, and survey results demonstrate relatively high consistency (Bill & Melinda Gates Foundation, 2012). In fact, the MET project reported that student surveys produce more consistent results than classroom observations or achievement gain measures (Kane & Staiger, 2012). Additionally, compared to classroom observations and teacher portfolios, student surveys can be relatively cheap and easy to administer. As a result, student surveys can be multiple times in a year, and can be administered early enough in the year to give teachers feedback so that their current students may benefit.

The major concern raised about student surveys is the worry that students are not objective raters of their teachers (Marsh, 1987; Theall & Franklin, 2001; Liaw & Goh, 2003). As MET researchers have acknowledged, "although most of the concern regarding bias has focused on the achievement gain measures, the non-test-based components could also be biased by the same unmeasured student traits that would cause bias with the test-based measures" (Kane, McCaffrey, Miller, & Staiger, 2013, p. 6). Following the definition used in Centra (2003, p. 498), bias exists when a "student, teacher, or course characteristic affects the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching, such as increased student learning." Popham (2013) outlines several sources of rater bias that may be a concern, including:

(a) *severity error* - a rater's predisposition to supply lower ratings independent of the content being rated, (b) *generosity error* - the opposite of severity error, where the rater will be more likely to use higher response categories, and (c) *central-tendency error* - the predisposition to use the middle rating category. A whole line of research regarding response styles has been developed to examine the impact of these types of rater bias on scores, but this work has not so far been applied to student ratings of instructional practice.

Another threat to the accuracy of ratings is the *halo effect*, which describes the tendency of a rater to rely on a single prominent dimension of the person being rated to distort the ratings of other dimensions. For example, if a teacher is particularly kind, a student may give the teacher positive ratings across various dimensions, including rigor and academic support, that may not be strongly related to kindness. Additionally, Popham (2013) warns of *lurking comparisons*, which is the tendency to use compare to other teachers when rating the characteristics of a given teacher. For example, a fifth-grade student might make implicit comparisons to their third and fourth grade math teachers when thinking about rating the kindness, rigor, and classroom management skills of his or her current teacher. This is a concern if teacher ratings are compared across a wide variety of contexts, because the examples of "quality teaching" that students are using in their comparisons may vary greatly.

1.4 Contributions of the current study

This study describes an innovative methodological approach for exploring the dimensionality and validity evidence to support the use of the Tripod student perception survey as a measure of teacher quality. This approach uses a multilevel extension of full-information item factor analysis models. Item factor analysis (IFA) models are widely-used in educational measurement research (Wirth & Edwards, 2007). These models are routinely used in calibration and scoring of large-scale educational assessments, as well as in the development of survey scales and observational protocols. IFA models have traditionally ignored the hierarchical structure of educational systems and treated all individuals in the sample as independent. Multilevel IFA models en-

able the data to be treated in an appropriate manner, instead of reducing the inferences to a single level. However, until recently, computational challenges estimating these highly-parameterized models have prevented the multilevel IFA models from entering mainstream educational research.

The aims of this dissertation are two-fold. First, I provide an introduction to the multilevel IFA modeling framework and demonstrate the flexibility and efficiency of these models in various educational settings. Unlike most previous multilevel IFA analyses in the literature (e.g., Fox & Glas, 2001; Kamata, 2001), I do not limit the analyses to unidimensional IFA models, but also explore multilevel correlated factors and testlet models. Secondly, I use multilevel IFA models to examine the dimensionality, reliability, and validity of the Tripod student survey.

The findings from this dissertation contribute to methodological and substantive bodies of work. This study provides an extension of the methodologies that have been developed in the context of large-scale educational surveys, such as National Assessment of Educational Progress (NAEP). In the current NAEP estimation approach, group-level scores are computed using a multi-step process that cycles through item calibration, latent regression, and score estimation. This study proposes a multilevel item factor analysis model that bridges the calibration, latent regression, and scoring into a single framework, and thanks to computational (Cai, 2010b) and software (Cai, 2015) advances, removes the need for multi-step estimation procedures.

I follow in the direction of other researchers that seek to provide a unified framework that combines multidimensional item factor models and nonlinear structural modeling in a multilevel setting (e.g., Rijmen, Jeon, von Davier, & Rabe-Hesketh, 2013; Rabe-Hesketh, Skrondal, & Pickles, 2004; Goldstein & Steele, 2005). While the discussion of including key conditioning covariates has been common-place with value-added models, it is has not been standard practice to include these variables in the analyses of the surveys of teacher practice. Within the multilevel IFA modeling framework described in this dissertation, it is possible for these statistical controls to go hand-in-hand with the dimensionality analysis, measurement modeling, validation efforts, and scoring/reporting. Additionally, models of higher dimensionality and complex-

ity than those previously investigated within the context of full-information multilevel item factor analysis are explored.

In addition, this dissertation contributes to the field of teacher evaluation by developing a validity argument for the widely used secondary Tripod survey. This work represents the first systematic review of the psychometric and validity properties of the secondary Tripod survey. Given the widespread use of the Tripod survey, it is important that there is sufficient evidence for its use in making high-stakes decisions about a teacher's effectiveness.

While others have previously examined the dimensionality and reliability of student ratings data (e.g., Abrami, d'Apollonia, & Rosenfield, 2007; Lüdtke, Robitzsch, Trautwein, & Kunter, 2009; Ferguson, 2010), this dissertation is an improvement on existing work due to four factors: (a) the use of multilevel item factor analysis allows for the estimation of a variety of different factor structures while accounting for the nesting of students in classrooms, (b) the use of the large and diverse sample of the Measures of Effective Teaching (MET) Project data is an improvement on previous studies that utilized a single school or district, (c) the MET data contain multiple other measures of teacher practice that can be used in validity analyses, and (d) the Tripod survey is the most widely-used measure of teacher practice, so understanding the factor structure of a survey that is currently used in summative teacher evaluation systems is of particular importance.

1.5 Chapter overview

This dissertation is organized as four separate research papers that address a range of topics related to multilevel IFA and student surveys of teacher practice. These papers will be stand-alone publications eventually, which necessitates some repetition of material across papers. Each paper is outlined briefly below.

The first paper examines applications of the multilevel IFA models across various educational settings. Simulations and data demonstrations in the context of large-scale educational assessment, student growth percentiles (SGPs), and student ratings

of instructional practice are analyzed as illustrations of the model's broad range of applicability.

The second paper proposes a novel approach for assessing model fit in multilevel IFA models. I first demonstrate theoretically and with a simulated data example that the multilevel IFA model can be re-parameterized into a single-level item bifactor model that is fit to the group-level data. By re-specifying the multilevel item factor analysis model to be estimated in a single-level context, the limited-information M_2 statistic (Maydeu-Olivares & Joe, 2006) is proposed for assessing model fit.

The third paper proposes a multilevel multidimensional plausible values approach for appropriately handling measurement error in the situation where latent teacher practice dimensions are used as predictor of student achievement in a hierarchical linear model. This approach borrows heavily from the multi-stage analytic procedures developed in the context of large-scale educational surveys (e.g., National Assessment of Educational Progress (NAEP)). The proposed approach consists of two stages: (a) specifying and imputing sets of plausible values from a multilevel item bifactor measurement model that consists of an overall teacher practice latent variable and seven latent domain-specific teacher practices, and (b) fitting a multilevel model to predict student achievement by the imputed teacher practice estimates. This modeling approach is illustrated using data from the Measures of Effective Teaching (MET) study, focusing on students within middle school English classrooms.

The last paper develops a validity argument for the secondary Tripod survey. I follow the validity framework developed by Kane (2006) to collect evidence to support two current uses of the Tripod survey—use in a summative high-stakes teacher evaluation system and use for feedback to teachers regarding strengths and weaknesses in their practice. The dimensionality, reliability, and validity of inferences based on the secondary Tripod survey are examined using student responses collected in English and math middle school classrooms by the MET Project.

Based on the results from these four lines of inquiry, I conclude with the lessons learned regarding the use of multilevel item factor analysis models for educational applications.

CHAPTER 2

Full-information multilevel item factor analysis with applications

2.1 Abstract

This paper demonstrates the utility of full-information multilevel item factor analysis models for a variety of different educational applications, including large-scale educational assessments and student surveys of teacher practice. Large-scale educational survey assessment researchers have developed a multi-stage analytic procedure called Latent Regression Item Response Theory that attempts to maximize the efficiency and precision of aggregated scores (von Davier & Sinharay, 2013). The focus of this paper is to develop the full-information multilevel item factor analysis model, which can be seen as a multilevel extension of the analytical models used in the estimation of large-scale assessment. This measurement framework allows for simultaneous estimation of item parameters, regression coefficients, and latent scores, and can accurately account for the multidimensional structure of data, nesting of students in classrooms, and background covariates. The Metropolis–Hastings Robbins–Monro (MH-RM) algorithm is used to estimate this model (Cai, 2010b). Empirical applications from a large-scale summative educational assessment, a student survey measure of teacher effectiveness, and estimation of student growth percentiles (SGP) for school-level inferences are provided as illustrations of the broad range of applicability of this modeling approach.

2.2 Introduction

In social and behavioral sciences, questionnaires or tests are often used to measure traits that are not directly observable. Item-level data from surveys and tests are gen-

erally categorical, with either dichotomous (yes/no, correct/incorrect) or polytomous (for example, strongly disagree, disagree, neither agree nor disagree, agree, strongly agree) item response formats. When items follow this response format and the latent (unobserved) variables are assumed to be continuous, item factor analysis (IFA) models are a flexible set of models that can be used to make inferences about the latent variables (Bock, Gibbons, & Muraki, 1988; Mislevy, 1986; Wirth & Edwards, 2007).

Recognizing that data collected in social and behavioral sciences frequently have a nested structure (e.g., students nested in classrooms and schools, patients nested within a treatment facility, individuals nested within a family), this paper examines applications of the full-information item factor analysis model to multilevel contexts. This work does not represent the first extension of item factor analysis into the multilevel context. As described in further detail below, the multilevel item factor analysis model has been developed over time in the item response theory tradition (e.g., Fox & Glas, 2001) as well as under the categorical confirmatory factor analysis framework (e.g., Asparouhov & Muthén, 2007; Muthén & Asparouhov, 2011). However, the full-information multilevel IFA framework described in this paper represents an improvement on prior multilevel IFA developments due to the fact that it allows for (a) flexible latent structures at both the within and between-level, including correlated factors, bifactor, and testlet factor structures, (b) the inclusion of observed individual- and group-level covariates at each level of the model, and (c) efficient estimation in existing commercial software (Cai, 2015).

This paper is organized as follows. First, the multilevel IFA model is described, and the estimation algorithm is outlined. Second, three motivating examples are outlined. The first example pertains to producing school-level subscores from a large-scale Common Core-aligned summative assessment. This example, which highlights the flexibility of the latent structure in the multilevel IFA framework, specifies a testlet structure (Wainer, Bradlow, & Wang, 2007) at the between-level to simultaneously estimate 28 school-level sub-domain scores. The second example draws from a large study of teacher effectiveness, and utilizes the latent regression functionality of the multilevel IFA model to examine the relationship between a latent teacher practice di-

mension and classroom compositional characteristics. The third example demonstrates how the multilevel IFA models can improve upon the inferences from standard methods for estimating aggregated student academic growth. This model demonstrates how multilevel IFA models are relevant to the educational policy conversations regarding the use of measures of student growth in teacher evaluation systems.

2.3 Multilevel item factor analysis (MLIFA) model

The multilevel item factor analysis model has been developed over time in the item response theory tradition. Fox and Glas (2001) and Fox (2010) have proposed a multilevel item response theory model that can accommodate dichotomous and polytomous items, and is estimated using Markov chain Monte Carlo (MCMC) methods. These developments also parallel a great deal of work that has been done in the structural equation modeling and nonlinear latent variable model traditions. Multilevel structural equation models (SEM) have emerged has a popular method for modeling latent variables with categorical manifest variables using an underlying variable approach (Muthén & Asparouhov, 2011; Skrondal & Rabe-Hesketh, 2005). Multilevel SEMs are generally estimated using limited-information estimation approaches (such as Weighted Least Squares estimation). Kamata (2001) described a hierarchical generalized linear model (HGLM) approach to multilevel item response theory model that permits investigation of the variation of students' performance across groups, and the interactive effect of person and group-characteristic variables. Additionally, frameworks uniting generalized linear mixed models, multilevel factor models, and item response models have been proposed by Rabe-Hesketh et al. (2004). The generalized linear latent and mixed models (GLLAMM) approach is estimated using adaptive quadrature.

The various modeling approaches described above all accommodate dichotomous and polytomous manifest variables, account for hierarchical nesting of individuals in groups, and allow for the inclusion of observed covariates in the latent factor model. However, high-dimensional full-information multilevel IFA models have not been widely used. The notable exception is multilevel IFA models estimated using weighted least squares (WLS) estimation (Asparouhov & Muthén, 2007), which have been previously

demonstrated (Muthén & Asparouhov, 2011, 2013). However, there are known concerns with the multi-stage limited-information estimation. Specifically, pairwise estimation of the polychoric correlation matrix can result in a correlation matrix that is not positive definite, and the treatment of missing data within WLS estimators is not ideal compared with full-information methods. MCMC estimation can be used for a general set of multilevel IRT models, although model specification and monitoring of convergence require some expertise. A quadrature-based full-information maximum likelihood (FIML) estimation approach will lead to analytically intractable integrals for high-dimensional IFA models.

With recent advances in FIML estimation algorithms (Cai, 2010a, 2010b) and software (Cai, 2015), it is now possible to estimate a high-dimensional item factor analysis model using FIML estimation in existing software that accounts for the complex sampling and multilevel data structures, incorporates the influence of background characteristics in the model, and allows for a wide variety of high-dimensional latent structures. The framework that allows for all of these conditions is multilevel IFA with latent regression, using the Metropolis–Hastings Robbins–Monro estimation algorithm (MH-RM Cai, 2010a).

2.3.1 Some notation

Let there be $p = 1, ...n$ items and $i = 1, ..., N_j$ individuals is nested in a group j, where $j = 1, ..., \mathrm{J}$ groups. Let the response from individual i in group j to item p be y_{pij}, where y_{pij} has K_p response categories. It can be assumed that y_{pij} takes integer values from $(0, \ldots, K_p - 1)$. Let the $n \times 1$ vector of item responses from respondent i in group j be $\mathbf{Y}_{ij} = (y_{1ij}, \ldots, y_{pij}, \ldots, y_{nij})$. The overall sample size is $\mathrm{N} = \sum_{j=1}^{\mathrm{J}} \mathrm{N}_j$.

In this model, the latent variables for individual i in group j are partitioned into two mutually exclusive parts: $\boldsymbol{\theta}_{ij} = \left(\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}\right)$, where $\boldsymbol{\vartheta}_j$ is the vector of group-level (level-2) latent variables, and $\boldsymbol{\eta}_{ij}$ is the vector of individual-level (level-1) latent variables. Within each of the latent variable vectors $\boldsymbol{\vartheta}_j$ and $\boldsymbol{\eta}_{ij}$, latent variables may be classified as (potentially correlated) primary latent dimensions or as specific dimensions that are independent conditional on the primary dimension(s).

The observed data (marginal) likelihood function may take the following form:

$$L(\boldsymbol{\gamma}) = \prod_{j=1}^{J} \int \prod_{i=1}^{N_j} \left[\int \prod_{p=1}^{n} P(Y_{pij} = y_{pij} | \boldsymbol{\theta}_{ij}) f\left(\boldsymbol{\eta}_{ij}\right) d\boldsymbol{\eta}_{ij} \right] f\left(\boldsymbol{\vartheta}_j\right) d\boldsymbol{\vartheta}_j, \tag{2.1}$$

where $\boldsymbol{\gamma}$ stands for the collection of freely estimated model parameters. The conditional distribution for an observed response y_{pij} is a multinomial distribution with trial size 1 in K_p cells, and the conditional density is

$$f_{\boldsymbol{\gamma}}(y_{pij} | \boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}) = \prod_{k=1}^{K_p - 1} P(y_{pij} = k | \boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij})^{\chi_k(y_{pij})}, \tag{2.2}$$

where $\chi_k\left(y_{pij}\right)$ is an indicator function that equals 1 when $y_{pij} = k$ and 0 otherwise.

Multilevel item factor analysis models contain two parts: a measurement model and a conditioning model.

2.3.2 Measurement model

The multilevel IFA model specifies the conditional probability for the response to item p with K_p categories from individual i in group j. Various standard IRT models could be used. Below, a few of the more well-known IRT models are described.

A Model for Dichotomous Response. The model below is an extension of the 3-parameter logistic (3PL) model. Let

$$P(y_{pij} = 1 | \boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}) = g_p + \frac{1 - g_p}{1 + \exp\left[-\left(c_p + \mathbf{a}_p^{B\prime} \boldsymbol{\vartheta}_j + \mathbf{a}_p^{\prime} \boldsymbol{\eta}_{ij}\right)\right]} \tag{2.3}$$

be the conditional probability of a correct response, where y_{pij} is a Bernoulli random variable representing the response to item p from individual i in group j, g_p is the lower asymptote (pseudo guessing) parameter, c_p is the item intercept, and $\mathbf{a}_p^{\prime}$ and $\mathbf{a}_p^{B\prime}$ are conformable vectors of level-1 and level-2 item slopes. The slope vectors can be constrained equal to each other across levels (representing cross-level invariance), but this is not a requirement for estimation. The model represents the response probability of a correct response ($y_{pij} = 1$) as a function of these item parameters and the latent

variables. The conditional probability of an incorrect response is $P(y_{pij} = 0|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}) = 1 - P(y_{pij} = 1|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij})$

A Model for Graded Response. Let y_{pij} be a polytomous item with K_p response categories, so that $y_{pij} \in (0, \ldots, K_p - 1)$. I use a multidimensional extension of the graded response model (Samejima, 1969). Let the cumulative category response probabilities be

$$\begin{aligned} P(y_{pij} \geq 1|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}) &= \frac{1}{1 + \exp\left[-\left(c_{p,k} + \mathbf{a}_p^{B'}\boldsymbol{\vartheta}_j + \mathbf{a}_p'\boldsymbol{\eta}_{ij}\right)\right]} \\ &\vdots \\ P(y_{pij} \geq K - 1|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}) &= \frac{1}{1 + \exp\left[-\left(c_{p,K-1} + \mathbf{a}_p^{B'}\boldsymbol{\vartheta}_j + \mathbf{a}_p'\boldsymbol{\eta}_{ij}\right)\right]} \end{aligned} \tag{2.4}$$

The category response probability is the difference between two adjacent cumulative probabilities

$$P(y_{pij} = k|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}) = P(y_{pij} \geq k|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}) - P(y_{pij} \geq k + 1|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}) \tag{2.5}$$

where $P(y_{pij} \geq 0|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij})$ is equal to 1 and $P(y_{pij} \geq K_p|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij})$ is zero. The item parameters $c_{p,1}, \ldots, c_{p,K-1}$ are a set of K-1 (strictly ordered) intercepts, and the item slopes are defined as in the dichotomous response model.

2.3.3 Conditioning model

Each of the latent variables within $\boldsymbol{\theta}_{ij} = \left(\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}\right)$ can be modeled as a linear function of person or group covariates using a latent regression model (Adams, Wilson, & Wu, 1997; Fox & Glas, 2001). For example, the latent regression model for a within-level latent variable could be

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \beta_{1j}\mathrm{X}_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\mathrm{W}_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \tag{2.6}$$

where β_{0j} is the intercept, and β_{1j} is a level-1 regression coefficient. The explanatory variables X_{ij} and W_j are included as individual and group-level covariates. The individual residuals e_{ij} are assumed to be normally distributed with zero means and variance σ^2, and the group-level random effects u_{0j} and u_{1j} are assumed to be normally distributed with zero means and variances τ_{00} and τ_{11}, and covariance τ_{01}.

Similarly, the latent regression model for a between-level latent variable could be

$$\vartheta_j = \beta_0 + \beta_1 \mathrm{W}_j + e_j \tag{2.7}$$

where β_0 is the level-2 intercept and β_1 is the level-2 slope coefficient.

2.3.4 Model estimation

The Metropolis–Hastings Robbins–Monro (MH-RM) algorithm is a data augmented Robbins-Monro type stochastic approximation algorithm using random imputations produced by a Metropolis–Hastings sampler (Cai, 2010a, 2010b). A major advantage of MH-RM is that its computational complexity is linear in the number of latent variables, whereas that of the EM algorithm with quadrature is exponential. Therefore, high-dimensional theoretical frameworks, which would not previously have been computationally feasible to estimate using the EM algorithm can be estimated with the MH–RM algorithm. Furthermore, this estimation algorithm allows for the concurrent estimation of item and latent regression parameters.

The Metropolis–Hastings Robbins–Monro algorithm is an iterative algorithm with three central steps in each cycle. The notation below includes covariates in the estimation, but this is not required for estimation. For cycle $b+1$, the algorithm proceeds as follows:

Imputation of $\boldsymbol{\theta}^{(b+1)}$. Given the parameter estimates from the previous cycle $\boldsymbol{\gamma}^b$ (which include item parameters and latent regression parameters), random samples of the latent variables $\boldsymbol{\theta}^{(b+1)}$ are imputed using the Metropolis–Hastings sampler with the posterior predictive distribution $P(\boldsymbol{\theta}_{ij}|\mathbf{Y}_{ij};\mathbf{X}_{ij};\boldsymbol{\gamma}^{(b)})$ of the missing data given the observed response $\mathbf{Y}_{ij}$, observed covariates $\mathbf{X}_{ij}$, and provisional

parameter values $\boldsymbol{\gamma}^{(b)}$. The complete datasets are formed as $(\boldsymbol{\theta}^{(b+1)}, \mathbf{Y}, \mathbf{X})$.

Approximation. Based on the imputed data, the complete data log-likelihood and its derivatives are evaluated so that the ascent direction for the item and latent density parameters can be determined.

Robbins–Monro update. Robbins–Monro stochastic approximation filters are applied when updating the estimates of item and latent density parameters.

The iterations are started from some initial parameter values $\boldsymbol{\gamma}^{(0)}$ and terminated when the estimates stabilize. The MH-RM algorithm is implemented in the software program flexMIRT® (Cai, 2015), which is used for calibration and scoring of the IRT models in this study.

2.3.5 Latent trait score estimation

Item factor analysis models are used to calculate a person's ability or trait level by first estimating the likelihood of the pattern of responses to the items, given the level on the underlying trait being measured by the scale. The data for an individual i in classroom j is a vector of responses of length n denoted by $\mathbf{Y}_{ij} = (y_{1ij}, y_{2ij}, ..., y_{pij}, ..., y_{nij})$ and the observed covariate vector $\mathbf{X}_{ij}$. The observed responses are assumed to be statistically independent given the latent traits ($\boldsymbol{\theta}_{ij} = \{\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}\}$).

I define $P_{kij}(\boldsymbol{\theta}) = P(y_{kij} = k|\boldsymbol{\theta}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\gamma})$ as the probability that the response y_{pij} of student i in classroom j to item p is the k-th category ($k = 0, \ldots, K_p - 1$), given the respondent's latent values $\boldsymbol{\theta}_{ij}$, the vector of observed covariate values $\mathbf{X}_{ij}$, and the pre-calibrated item parameters, latent distribution, and regression parameters within the parameter vector $\boldsymbol{\gamma}$. Therefore, I can define the probability of the set of item responses for a student given a particular $\boldsymbol{\theta}_{ij}$-vector by the likelihood function

$$L(\mathbf{Y}_{ij}|\boldsymbol{\theta}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\gamma}) = \prod_{p=1}^{n} \prod_{k=0}^{K_p-1} P_{pij}(\boldsymbol{\theta}_{ij})^{\chi_k}, \tag{2.8}$$

which is the product of the individual item trace lines P_{pij}. In the likelihood formula, χ_k is a random variable defined in Equation 2.2. Each trace line models the probability

of a response to the item p conditional on the underlying trait vector $\boldsymbol{\theta}_{ij}$. It is typically assumed that response probabilities are conditionally independent of background variables $\mathbf{X}_{ij}$ given $\boldsymbol{\theta}_{ij}$, which is to say

$$L(\mathbf{Y}_{ij}|\boldsymbol{\theta}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\gamma}) = L(\mathbf{Y}_{ij}|\boldsymbol{\theta}_{ij}, \boldsymbol{\gamma}). \tag{2.9}$$

The maximum likelihood (ML) estimator of trait values $\hat{\boldsymbol{\theta}}_{ML}$ is defined as the values of $\boldsymbol{\theta}$ that maximize the likelihood function given in Equation 2.8.

Estimation of latent trait values $\boldsymbol{\theta}$ can also be obtained through Bayesian methods. Using estimated item parameters $\boldsymbol{\gamma}$, inference for $\boldsymbol{\theta}$ is based on the following posterior:

$$f(\boldsymbol{\theta}|\mathbf{Y}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\gamma}) = \frac{L(\mathbf{Y}_{ij}|\boldsymbol{\theta}, \boldsymbol{\gamma}) f(\boldsymbol{\theta}|\mathbf{X}_{ij})}{\int_{\boldsymbol{\theta}} L(\mathbf{Y}_{ij}|\boldsymbol{\theta}, \boldsymbol{\gamma}) f(\boldsymbol{\theta}|\mathbf{X}_{ij}) \mathrm{d}\boldsymbol{\theta}}, \tag{2.10}$$

where $f(\boldsymbol{\theta}|\mathbf{X}_{ij})$ is the prior distribution of $\boldsymbol{\theta}$ given the observed (person) background covariates $\mathbf{X}_{ij}$. The Expected A Posteriori (EAP) estimate of an respondent's trait level is a commonly used Bayesian estimator in IRT (Bock & Mislevy, 1982). The EAP estimator is calculated by taking the expectation over posterior distribution in Equation 2.10

$$\hat{\boldsymbol{\theta}}_{ij} = \int_{\boldsymbol{\theta}} \boldsymbol{\theta} f(\boldsymbol{\theta}|\mathbf{Y}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\gamma}) d\boldsymbol{\theta} = \frac{1}{f(\mathbf{Y}_{ij}|\mathbf{X}_{ij}, \boldsymbol{\gamma})} \int_{\boldsymbol{\theta}} \boldsymbol{\theta} L(\mathbf{Y}_{ij}|\boldsymbol{\theta}, \boldsymbol{\gamma}) f(\boldsymbol{\theta}|\mathbf{X}_{ij}) d\boldsymbol{\theta}, \tag{2.11}$$

with the standard error of measurement given by the square root of posterior variance

$$V(\hat{\boldsymbol{\theta}}_{ij}) = \int_{\boldsymbol{\theta}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{ij})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{ij})' f(\boldsymbol{\theta}|\mathbf{Y}_{ij}, \mathbf{X}_{ij}, \boldsymbol{\gamma}) d\boldsymbol{\theta}. \tag{2.12}$$

2.4 Educational applications of multilevel item factor analysis

2.4.1 Subscore recovery in large-scale summative assessments

With the adoption of new Common Core Standards, a new generation of large-scale summative assessments moved toward a multi-layered test design, where content, practice, and cognition-related facets each contribute to an item or task's role in reflecting

the domain being measured (National Governors Association, 2014). Another distinguishing feature of these new assessments is the wide-spread use of technology to administer the assessments, and the use of Item Response Theory in test construction, administration (e.g., computer adaptive testing (CAT)), and scoring.

The Smarter Balanced Assessment Consortium is a consortium of 15 states that has developed standardized summative assessments for Grades 3-8 and 11 that are aligned with the Common Core State Standards (Smarter Balanced Assessment Consortium, 2016). The Smarter Balanced Assessment Consortium test specifications are highly structured hierarchically, with items nested within each of the Assessment Targets under each of the claims. Assessment Claims are broad evidence-based statements about what students know and can do as demonstrated by their performance on the assessments. Each Assessment Claim is accompanied by a set of Assessment Targets, which map the standards in the Common Core State Standards onto assessment evidence that is required to support the content categories and Claims (Smarter Balanced Assessment Consortium, 2013).

Substantial time and money has been invested in the new state accountability assessments, and educators and parents prefer to receive both domain and sub-domain information from these assessments to explain the student's performance in English and Mathematics, as well as to evaluate the effects of teaching practices in classroom. However, given the test blueprint structure and time limitations, each student only typically receives zero to three items per Target sub-domain, which is generally not a sufficient length to produce reliable student subscores. Using multilevel IFA models, I describe a method for producing school-level subscores that could provide meaningful feedback about student performance, and is also feasible within the constraints of the operational testing structure. Target subscores that are reported at the school-level are also useful because a school's pattern of Target scores may provide useful information for investigating the implementation of the Common Core State Standards. Level of implementation of standards is likely to vary across schools and therefore these school-level Target scores allow for potential comparisons across schools and across years in the degree to which the Common Core standards have been successfully implemented

in schools.

In this application, I focus on the Grade 4 Mathematics Smarter Balanced end-of-year assessment. Table 2.1 provides an overview of the Assessment Claims and Assessment Targets (generally referred to simply as Targets) that are specified within Grade 4 Mathematics. These Targets are tied to specific Common Core standards, and represent more granular skills needed to be considered proficient in Mathematics in fourth grade. The test blueprint specifies that students receive between 37-40 total items in the summative assessment. As seen in Table 2.1, there are 29 Targets. Therefore, it is clear that student-level subscores will not be very reliable, as they would be based on an item or two at most. Therefore, the purpose of this example is to investigate whether using a multilevel IFA model to produce school-level subscores results in the reliable estimation of Target scores.

Given that the test blueprint assumes the Mathematics assessment is assessing general math proficiency in addition to more granular skills, a multilevel testlet factor analysis model is proposed for this application. An example path diagram is displayed in Figure 2.1. In this model, each item loads on the primary (domain-general) school-level dimension (ϑ_j^G), a primary (domain-general) student-level dimension (η_{ij}; which can be seen as a deviation from the school general dimension), as well as a single school-level specific dimension ($\vartheta_j^{T_1}$).

For a dichotomous 2PL item that measures Target 1, the probability of a correct response would be

$$P(y_{pij} = 1|\vartheta_j^G, \vartheta_j^{T_1}, \eta_{ij}) = \frac{1}{1 + \exp\left[-\left(c_p + a_p(\vartheta_j^G + \vartheta_j^{T_1} + \eta_{ij})\right)\right]}. \tag{2.13}$$

Field test item calibrations provide the location and general estimates of the discriminability of the items relative to one latent continuum that represents the general domain. Thus the task of inferring student achievement at the Target level is reduced to the task of variance decomposition, followed by empirical Bayes (augmented or regressed) estimates. I submit that the underlying total variance may be decomposed in the following way:

Variance of a domain = Domain-general student variance + domain-general school-level variance + Target-specific variance

This model is demonstrated using simulated data based upon the Grade 4 Mathematics Smarter Balanced Assessment Consortium blueprint and field test data. There are in total 916 CAT items in the fourth grade math summative assessment pool, representing 29 Targets listed in Table 2.1. Therefore, the multilevel testlet model contains 31 dimensions, one student-level general dimension, one school-level general dimensions, and 29 school-level specific factors. True latent proficiency scores are generated for N = 10,000 students, who are nested in J = 100 schools, following

$$\begin{aligned} \vartheta_j^G &\sim N\left(0, \tau_G^2\right), \\ \vartheta_j^{T_1} &\sim N\left(0, \tau_{T_1}^2\right), \\ &\vdots \\ \vartheta_j^{T_m} &\sim N\left(0, \tau_{T_m}^2\right), \\ \eta_{ij} &\sim N\left(0, \sigma_G^2\right), \end{aligned} \tag{2.14}$$

where $\sigma_G^2 = 1$ represents the student domain-general variance, $\tau_{T_m}^2$ represents the school-level Target-specific variance for the mth Target (true values range from .03 to .11, and are printed in Table 2.2), and $\tau_G^2 = 1$ represents school-level domain-general variance. These latent proficiencies were fed into a computerized adaptive testing (CAT) algorithm programmed in R (R Core Team, 2012) that was constructed to match the Smarter Balanced blueprint specifications and utilized item parameter estimates from prior field tests (National Center for Research on Evaluation, Standards, and Student Testing, 2015). Each simulee received between approximately 40 multiple choice and performance task items, with latent proficiency estimates updated after each item response. The response simulation was conducted on a standard 64-bit Windows desktop computer with an Intel Core 7 CPU at 3.60 GHz and 16 GB of RAM.

After all of the responses were collected, a final scoring dataset was created with rows for all of the respondents and columns representing all of the 4th grade mathematics items. For each respondent, any non-observed items were coded as missing. A

multilevel item testlet analysis model was specified in flexMIRT (Cai, 2015) fixing all of the item parameters (2PL and generalized partial credit (GPC)) to the estimates from the field test, and freeing all of the latent variance parameters. The 31-dimensional multilevel IFA model took 5.8 hours to converge. EAP scores were estimated from this model for student general proficiency, school general proficiency, and school-level Target proficiency.

Three outcomes were examined in this study. First, the average bias of the school-level score estimates were calculated. Given a school's true score, ϑ_j, and final score estimate, $\hat{\vartheta}_j$, average bias in the score estimates is defined as

$$\text{bias} = J^{-1} \sum_{j=1}^{J} \left(\vartheta_j - \hat{\vartheta}_j \right), \tag{2.15}$$

and the error variance of the estimated bias is

$$\text{var(bias)} = \frac{1}{J(J-1)} \sum_{j=1}^{J} \left(\vartheta_j - \overline{\hat{\vartheta}} \right)^2, \tag{2.16}$$

where $\overline{\hat{\vartheta}}$ is the average of the $\hat{\vartheta}_j$ and J denotes the number of schools (J =100). Secondly, the variance recovery for the school general dimension (ϑ_j^G) and the 29 specific factors ($\vartheta_j^{T_1}, ..., \vartheta_j^{T_{29}}$) are examined. Lastly, a marginal reliability index ρ_ϑ summarizes the reliability of a measure as the proportion of variance in the observed score that is due to the true score

$$\rho_\vartheta = \frac{\sigma_\vartheta^2 - \sigma_e^2(\vartheta)}{\sigma_\vartheta^2} = 1 - \frac{\sigma_e^2(\vartheta)}{\sigma_\vartheta^2} \tag{2.17}$$

where σ_ϑ^2 is the prior value of the variance of ϑ, and $\sigma_e^2(\vartheta)$ is the marginal or average error variance of ϑ. The error variance $\sigma_e^2(\vartheta)$ can be calculated as an average over a random sample of individuals from the population distribution

$$\overline{\text{SE}}^2 = \frac{1}{\text{J}} \sum_{j=1}^{\text{J}} \text{SE}^2(\vartheta_j) \tag{2.18}$$

where $\text{SE}^2(\vartheta_j)$ is the squared standard error for the jth school.

Table 2.2 displays the average number of item responses per dimension that were

observed in each school, measures of ϑ recovery, the true and estimated latent variances, and the marginal reliability estimates per Target. Due to constraints in the test blueprint, the number of item responses per school used to calculate each Target score ranged widely, from nine item responses at the lowest and 519 item responses at the highest. The school-level ϑ_j scores were well-recovered. Bias in the estimated score was small (ranging from -0.03 to 0.04), with the estimated bias of most of the latent dimensions equal to zero. The generating variances were well estimated for most of the Target dimensions, but a few of the variance estimates were inflated. Lastly, the marginal reliability of the Target dimensions were low for Targets that were measured by less than 100 responses per school, and were greater than .70 for Targets that were measured by larger sets of item responses.

2.4.2 Student surveys of teacher effectiveness

For this application, I focus on the secondary Tripod survey as a measure of teachers' instructional practice. Student surveys of instructional practice and classroom environment have recently gained prominence among researchers and policymakers as an inexpensive way to get feedback on aspects of the teacher's instruction and classroom environment. Seven states now mandate that student surveys are required as a component of teacher evaluations, while 26 other states allow for the use of students surveys in teacher evaluations (Doherty & Jacobs, 2015). The Tripod survey, a widely-used student perceptions measure, asks students to rate seven dimensions of classroom instruction as they experienced it in a given classroom (Ferguson, 2010). The stated intention of the Tripod survey is to be both a diagnostic and professional development tool. Specifically, the survey can be used to provide information that will guide setting professional development and coaching goals (Tripod Project, 2016).

While these surveys are now widely used, it is not well known how perceptions of teachers differs across different classroom and school contexts. Multilevel item factor analysis can be used to measure whether background characteristics and traits of students and teachers are negatively or positively related to student perceptions' of instructional practice. Observed differences in student perceptions due to classroom

compositional characteristics may reflect true differences in teaching quality that stem from systematic selection processes in the assignments of students to teachers. For instance, if more senior teachers tend to be better teachers and also use their seniority to teach more honors courses, then the data might show an association between teacher practice and student prior achievement. Alternatively, differences in student perceptions due to background characteristics may reflect ratings bias, which would occur if a characteristic of students or teachers affected the evaluations made, but were unrelated to the true teacher practice. This may occur if male students, for example, always rate their female teachers lower than male teachers, even if male and female teachers teach equally well on average. It is difficult to disentangle true differences in teaching practice from the existence of ratings bias, and therefore the purpose of this study is to examine whether there are observed characteristics of classrooms that are significantly associated with negative or positive ratings of teacher practice.

This example uses data from the Measures of Effective Teaching (MET) project, which is the largest study of classroom teaching ever conducted in the United States. Data were collected on a variety of measures of teacher quality over two academic school years, 2009-2010 and 2010-2011, within six large school districts in the United States. A sample of approximately 3,000 teacher volunteers was recruited from six urban districts: Charlotte-Mecklenburg (North Carolina) Schools, Dallas (Texas) Independent School District, Denver (Colorado) Public Schools, Hillsborough County (Florida) Public Schools, Memphis (Tennessee) City Schools, and the New York City (New York) Department of Education (White & Rowan, 2013).

For this study, 19,046 students in 1,071 middle school English/Language Arts classrooms during the 2009-10 school year are included. There are total 36 items in the secondary survey related to the dimensions of classroom instruction, and students are asked to rate teacher practices using a Likert-type response options with a 5-point scale (Totally Untrue; Mostly Untrue; Somewhat; Mostly True; Totally True). The Tripod items are organized under seven constructs, called the 7Cs$^{\text{TM}}$: Care, Control, Clarify, Challenge, Captivate, Confer, and Consolidate. While there are seven theoretical constructs, the scores from this measure have been found to be highly inter-correlated (see

Raudenbush & Jean, 2014), and scores are often just reported as an overall, composite measure of teacher practice (e.g., Kane et al., 2013).

A multilevel unidimensional model was fit to the secondary Tripod survey, with all items loading on an overall classroom-level "teacher practice" factor and a student-level latent factor representing the student deviation from the mean. The path diagram for this model is provided in Figure 2.2. The item slopes were fixed to be equal across levels, representing cross-level measurement invariance (e.g., there is invariance in the measurement structure across the individual-level and the between-classroom level (Bliese, 2000)). The classroom-level latent dimension was regressed on the percentage of students in the class who are in a special education program, class average of a student-reported measure of effort put forth in class, the percentage of minority students in class, and average prior year test score. The percentage minority and percentage special education range from 0-1, and the class effort and prior year test scores are scaled to follow a standard normal distribution.

Tripod item wording and estimated item parameters are reported in Table 2.3, and the latent regression parameters are reported in Table 2.4. The class-level variance was estimated to be 0.20, relative to the student-level variance which is fixed to 1.0. The intraclass correlation (ICC) of the teacher practice dimension was estimated to be 0.17, implying that most of the variance in the student perceptions of teacher practice was between students within a classroom. This percentage is consistent with previous analyses of student surveys (e.g., Phillips, Yamashiro, Schaaf, & Schweig, 2011). The latent regression estimates demonstrate that classrooms with a high percentage of special education students and higher levels of student-reported effort were more likely to have positive ratings of teachers. However, ratings of teachers did not vary greatly by percentage of minority students or students' prior year test scores in English/Language Arts.

These results demonstrate that there are student characteristics that are significantly related to student perceptions of teacher practice as measured by the Tripod survey. However, it is not clear from these results the directionality of the findings. In the example of the student effort variable, it may be that that students who put in

more effort also are the types of kids who rate their teachers higher. It also may be that better teachers sort into teaching students that exert higher effort in class. Further study is needed to explain the observed relationships and determine the degree to which differences in classroom composition variables result in biased teacher scores.

2.4.3 Student growth percentiles (SGPs)

Student Growth Percentiles (SGP; Betebenner, 2009) locate a student's current score in a conditional distribution depending on the student's past scores. Numerous states use aggregates of SGP estimates (e.g., means) to evaluate teachers, with the desired inference being that higher aggregate scores reflect higher levels of teacher effectiveness. However, an alternative to aggregated SGPs may be defined, and estimated using multilevel IFA. This measure, a Cluster Growth Percentile (CGP), is defined analogously to an SGP. Whereas an SGP depends on a student's scores, a CGP depends on a classroom's aggregated scores ($\boldsymbol{\vartheta}_j$). In this section, the CGP framework is presented, and CGPs are compared with aggregated SGPs using simulated data examples.

To facilitate the presentation, only one prior year is considered. Let Z_p and Z_c be prior and current scores, respectively. Then, a growth percentile is defined as

$$G\left(Z_c,\ Z_p\right) = \int_{-\infty}^{Z_C} p(t|Z_p)dt, \tag{2.19}$$

where $p(Z_c|Z_p)$ is determined by the joint distribution $p(Z_c, Z_p)$. Following Lockwood and Castellano (2015), an SGP is defined as

$$S = G\left(\theta_c,\ \theta_p\right), \tag{2.20}$$

where θ_c and θ_p are latent student scores, and S depends on $p(\theta_c, \theta_p)$, the joint distribution of the latent scores. By construction, S is uniformly distributed.

For teacher-evaluation, S may be aggregated. Following Shang, VanIwaarden, and

Betebenner (2015), I use the mean,

$$M_j = \frac{1}{N_j} \sum_{i=1}^{N_j} S_{ij}, \tag{2.21}$$

where S_{ij} is the SGP of the ith student in classroom j, and N_j is the class size. Unlike S, in general, the distribution of M is unknown. The CGP uses the decomposition

$$\theta_{ij} = \vartheta_j + \eta_{ij}, \tag{2.22}$$

where ϑ_j is the latent classroom mean, and η_{ij} is the student deviation. By definition, $\vartheta_j = N_j^{-1} \sum_{j=1}^{N_j} \theta_{ij}$, and $\sum_{j=1}^{N_J} \eta_{ij} = 0$. Let $p(\vartheta_c, \vartheta_p)$ be the distribution of latent means. Then, the CGP is

$$C_j = G\left(\vartheta_{cj},\ \vartheta_{pj}\right). \tag{2.23}$$

Like S, C is uniformly distributed. To summarize the distinction between M and C, M is the mean of growth percentiles, and C is the growth percentile of means.

As with S, C may be estimated using either quantile regression (QR) or IFA. Using QR, estimates of $\boldsymbol{\theta} = (\theta_c, \theta_p)'$ are first obtained. Then, these estimates are averaged to obtain an estimate of the latent mean $\boldsymbol{\vartheta} = (\vartheta_c, \vartheta_p)'$. Finally, these estimates may be analyzed by QR to yield estimates of C.

Estimates of C may also be obtained with a multilevel IFA model. The latent scores for Levels 1 and 2 are $\boldsymbol{\eta} = (\eta_c, \eta_p)'$ and $\boldsymbol{\vartheta} = (\vartheta_c, \vartheta_p)'$, respectively. The path diagram for this model is presented in Figure 2.3. Of particular interest in this framework is the distribution of the group latent variables, $p(\vartheta_c, \vartheta_p)$, as this is used in the definition of C (see Equation 2.23).

One way to identify the model is as follows:

1. For the item parameters, constrain the slopes (a^B and a) across levels to be equal.

2. For the Level 1 latent score distribution, estimate the covariance. Fix the variances (e.g., to 1.0) to identify the model.

3. For the Level 2 latent score distribution, estimate the variances and covariance.

For the latent scores (at both levels), bivariate normal distributions are specified. Following calibration, estimates of C may be obtained directly. In particular, the EAP estimate, $\hat{C}$, may be calculated in an entirely analogous manner to the EAP estimate of S (Lockwood & Castellano, 2015; Monroe & Cai, 2015).

A single dataset for each of 16 conditions was simulated to evaluate the proposed framework. For all datasets, the multilevel IFA model described above was used to simulate the true latent values (e.g., θ, M, C). Item responses were generated using a 3PL model, as in Equation 2.3. For each year, the marginal reliability of the test was 0.90. The manipulated factors were: number of Level-1 units per Level-2 unit (20 or 30), number of Level-2 units (250 or 500), latent distribution correlations for Levels 1 and 2 (0.7 and 0.75; or 0.8 and 0.85), and intraclass correlations for the prior and current year dimensions (0.2 and 0.3; or 0.25 and 0.35).

A multilevel IFA model was used to obtain $\hat{C}$. Table 2.5 presents some key statistics: correlations between (true) M and C; correlations between C and $\hat{C}$; and the marginal reliability of C (Monroe & Cai, 2015). Key findings include: 1) M and C are highly correlated, but clearly not identical; 2) the correlations between C and $\hat{C}$ are high, indicating the rank-ordering of C can be well-recovered using the multilevel IFA model; and 3) the marginal reliabilities range from 0.68 to 0.88, indicating that C is arguably sufficiently reliable for operational use under certain circumstances.

2.5 Discussion

The multilevel item factor analysis model estimated by full-information maximum likelihood is described in this paper. The analysis of several real and simulated data examples in multiple policy-relevant educational domains, including large-scale assessments and teacher evaluation measures, serve as illustrations of the wide ranging applicability of the multilevel IFA modeling framework. The multilevel IFA model is flexible enough to represent a variety of structures commonly found in educational measurement, including the correlated factors structure used in the SGP application and the testlet factor model used in the large-scale assessment application. The Metropolis–Hastings

Robbins–Monro algorithm implemented in the flexMIRT item response theory (IRT) software (Cai, 2015) allows for the estimation of even high-dimensional multilevel item factor models under reasonable computational time.

Additionally, the multilevel IFA model is flexible in its ability to incorporate individual and group covariates into the model. Standard practice in IFA modeling has been that no information beyond the items themselves and distributional assumptions about the latent variable (usually assumed to be standard normal) are used in model calibration as well in scaled score estimation. However, as seen in the Tripod survey example, inclusion of covariates in the IFA model allows for a better understanding of how the latent trait of interest differs across contexts. Furthermore, accounting for person and group covariates in the estimation of the individual latent estimates leads to factor score estimates that are more precise.

The last application considered was cluster growth percentiles. Student growth percentiles (SGP) are now widely used in the United States to gauge the academic progress of both individual students, and are increasingly being aggregated to measure the growth of students in classrooms for teacher evaluative purposes (Lockwood & Castellano, 2015). This paper uses a correlated traits multilevel item factor analysis as a means to estimate classroom-level student growth EAP scores, which are referred to as Cluster Growth Percentiles (CGP). The CGP is an alternative to aggregated SGPs and the measures differ in important ways. First, the CGP is structurally analogous to the SGP, meaning that techniques developed for SGPs (e.g., SIMEX bias-correction, Shang et al., 2015) may be applied to CGPs. Second, since the CGP is uniformly distributed, estimates and their standard errors are easier to interpret. I demonstrate the reliability of cluster growth scores is dependent on sample size and the correlations of latent traits across years. For states implementing a student growth measure in teacher evaluation systems, careful thought should be given when selecting among CGPs and aggregated SGPs.

Table 2.1: Smarter Balanced Assessment Consortium Mathematics blueprint for grade 4

Claim	Content Category	Assessment Targets
Concepts and Procedures	Priority Cluster	A. Use the four operations with whole numbers to solve problems. E. Use place value understanding and properties of operations to perform multi-digit arithmetic. F. Extend understanding of fraction equivalence and ordering. G. Build fractions from unit fractions by applying and extending previous understandings D. Generalize place value understanding for multi-digit whole numbers. H. Understand decimal notation for fractions, and compare decimal fractions.
	Supporting Cluster	I. Solve problems involving measurement and conversion of measurements from a larger unit to a smaller unit. K. Geometric measurement: understand concepts of angle and measure angles. B. Gain familiarity with factors and multiples. J. Represent and interpret data. L. Draw and identify lines and angles, and classify shapes by properties of their lines and angles.
Problem Solving and Modeling and Data analysis	Problem Solving	A. Apply mathematics to solve well-posed problems arising in everyday life, society, and the workplace. B. Select and use appropriate tools strategically. C. Interpret results in the context of a situation. D. Identify important quantities in a practical situation and map their relationships.
		A. Apply mathematics to solve well-posed problems arising in everyday life, society, and the workplace. D. Interpret results in the context of a situation. B. Construct, autonomously, chains of reasoning to justify mathematical models used, interpretations made, and solutions proposed for a complex problem.

	Modeling and Data Analysis	E. Analyze the adequacy of and make improvements to an existing model or develop a mathematical model of a real phenomenon. C. State logical assumptions being used. F. Identify important quantities in a practical situation and map their relationships. G. Identify, analyze, and synthesize relevant external resources to pose or solve problems.
Communicating Reasoning		A. Test propositions or conjectures with specific examples. D. Use the technique of breaking an argument into cases. B. Construct, autonomously, chains of reasoning that will justify or refute propositions or conjectures. E. Distinguish correct logic or reasoning from that which is flawed, and–if there is a flaw in the argument–explain what it is. C. State logical assumptions being used. F. Base arguments on concrete referents such as objects, drawings, diagrams, and actions.

Table 2.2: Results for Smarter Balanced Assessment Consortium simulated data analysis

	Average number of item responses per school	θ recovery		τ^2 recovery		Marginal Reliability
		Bias	Var (Bias)	True Var.	Est. Var.	
School General Dim.	3915	-0.02	0.00	0.11	0.12	0.88
Claim 1. Target A	35	-0.02	0.00	0.06	0.05	0.53
Claim 1. Target E	54	0.00	0.00	0.05	0.04	0.71
Claim 1. Target F	11	-0.03	0.00	0.03	0.03	0.36
Claim 1. Target G	200	0.01	0.00	0.04	0.03	0.79
Claim 1. Target D	346	0.01	0.00	0.07	0.06	0.78
Claim 1. Target H	519	0.01	0.00	0.07	0.11	0.83
Claim 1. Target I	300	0.00	0.00	0.08	0.07	0.76
Claim 1. Target K	100	0.02	0.00	0.07	0.07	0.75
Claim 1. Target B	154	0.00	0.00	0.09	0.11	0.81
Claim 1. Target C	35	0.01	0.00	0.09	0.17	0.51
Claim 1. Target J	146	0.01	0.00	0.08	0.05	0.70
Claim 1. Target L	100	0.01	0.00	0.08	0.09	0.63
Claim 2. Target A	351	0.01	0.00	0.03	0.02	0.70
Claim 2. Target B	9	-0.02	0.00	0.09	0.14	0.33
Claim 2. Target C	77	0.02	0.00	0.08	0.08	0.71
Claim 2. Target D	51	0.04	0.00	0.09	0.10	0.47
Claim 4. Target A	195	0.00	0.00	0.04	0.05	0.82
Claim 4. Target B	167	0.00	0.00	0.08	0.06	0.72
Claim 4. Target C	138	0.00	0.00	0.09	0.12	0.70
Claim 4. Target D	126	0.01	0.00	0.07	0.07	0.76
Claim 4. Target E	259	0.00	0.00	0.04	0.04	0.79
Claim 4. Target F	72	-0.01	0.00	0.09	0.07	0.64
Claim 4. Target G	170	0.00	0.00	0.03	0.04	0.72
Claim 3. Target A	36	0.00	0.00	0.07	0.07	0.59
Claim 3. Target B	36	0.00	0.00	0.05	0.04	0.63
Claim 3. Target C	36	-0.02	0.00	0.05	0.04	0.51
Claim 3. Target D	96	-0.01	0.00	0.06	0.06	0.55
Claim 3. Target E	85	0.03	0.00	0.05	0.04	0.73
Claim 3. Target F	11	-0.01	0.00	0.08	0.08	0.15

Table 2.3: Results from the Tripod survey calibration

Item	a1	a2	c1	c2	c3	c4
My teacher in this class makes me feel that s/he really cares about me.	1.81(.02)	1.81(.02)	2.91(.03)	1.66(.02)	-0.39(.02)	-2.11(.03)
My teacher seems to know if something is bothering me.	1.29(.02)	1.29(.02)	1.55(.02)	0.38(.02)	-1.29(.02)	-2.77(.03)
My teacher really tries to understand how students feel about things.	1.93(.02)	1.93(.02)	2.96(.04)	1.59(.03)	-0.7(.02)	-2.85(.03)
Student behavior in this class is under control.	0.85(.01)	0.85(.01)	1.99(.02)	0.95(.02)	-0.5(.02)	-2.04(.02)
I hate the way that students behave in this class.	0.29(.01)	0.29(.01)	1.95(.02)	1.13(.02)	0.01(.02)	-1.01(.02)
Student behavior in this class makes the teacher angry.	0.56(.01)	0.56(.01)	1.34(.02)	0.37(.02)	-0.95(.02)	-2.15(.02)
Student behavior in this class is a problem.	0.54(.01)	0.54(.01)	1.83(.02)	0.87(.02)	-0.41(.02)	-1.71(.02)
My classmates behave the way my teacher wants them to.	1.02(.01)	1.02(.01)	1.81(.02)	0.64(.02)	-1.09(.02)	-2.87(.03)
Students in this class treat the teacher with respect.	1.09(.01)	1.09(.01)	2.7(.03)	1.52(.02)	-0.21(.02)	-2.11(.02)
Our class stays busy and does not waste time.	0.97(.01)	0.97(.01)	2.47(.03)	1.23(.02)	-0.49(.02)	-2.27(.02)
If you don't understand something, my teacher explains it another way.	1.80(.02)	1.80(.02)	4.15(.05)	2.75(.03)	0.61(.02)	-1.61(.02)
My teacher knows when the class understands, and when we do not.	1.42(.02)	1.42(.02)	3.44(.04)	2.14(.03)	0.05(.02)	-1.94(.02)
When s/he is teaching us, my teacher thinks we understand even when we don't.	0.76(.01)	0.76(.01)	2.45(.03)	1.37(.02)	-0.06(.02)	-1.49(.02)
My teacher has several good ways to explain each topic that we cover in this class.	2.06(.02)	2.06(.02)	4.42(.05)	2.78(.03)	0.21(.02)	-2.26(.03)
My teacher explains difficult things clearly.	1.98(.02)	1.98(.02)	3.96(.05)	2.54(.03)	0.04(.02)	-2.34(.03)
My teacher asks questions to be sure we are following along when s/he is teaching.	1.21(.02)	1.21(.02)	3.96(.05)	2.88(.03)	1.30(.02)	-0.46(.02)
My teacher asks students to explain more about answers they give.	0.99(.02)	0.99(.02)	3.83(.05)	2.57(.03)	0.76(.02)	-1.1(.02)

Item	a1	a2	c1	c2	c3	c4
In this class, my teacher accepts nothing less than our full effort.	1.21(.02)	1.21(.02)	3.52(.04)	2.32(.03)	0.38(.02)	-1.27(.02)
My teacher doesn't let people give up when the work gets hard.	1.53(.02)	1.53(.02)	3.58(.04)	2.41(.03)	0.53(.02)	-1.36(.02)
My teacher wants us to use our thinking skills, not just memorize things.	1.38(.02)	1.38(.02)	3.86(.05)	2.78(.03)	0.79(.02)	-1.15(.02)
My teacher wants me to explain my answers–why I think what I think.	1.30(.02)	1.30(.02)	3.90(.05)	2.65(.03)	0.64(.02)	-1.16(.02)
In this class, we learn a lot almost every day.	1.65(.02)	1.65(.02)	3.62(.04)	2.05(.03)	-0.09(.02)	-2.22(.03)
In this class, we learn to correct our mistakes.	1.82(.02)	1.82(.02)	4.14(.05)	2.78(.03)	0.47(.02)	-1.75(.02)
This class does not keep my attention–I get bored.	1.12(.02)	1.12(.02)	1.83(.02)	0.81(.02)	-0.57(.02)	-1.93(.02)
My teacher makes learning enjoyable.	2.06(.02)	2.06(.02)	2.87(.03)	1.57(.03)	-0.72(.02)	-2.6(.03)
My teacher makes lessons interesting.	2.14(.02)	2.14(.02)	3.02(.04)	1.59(.03)	-0.81(.02)	-2.89(.03)
I like the ways we learn in this class.	1.62(.02)	1.62(.02)	3.85(.05)	2.75(.03)	0.11(.02)	-2.32(.03)
My teacher wants us to share our thoughts.	1.07(.02)	1.07(.02)	3.08(.04)	2.05(.03)	0.33(.02)	-1.29(.02)
Students get to decide how activities are done in this class.	0.71(.01)	0.71(.01)	0.90(.02)	-0.33(.02)	-2.35(.02)	-4.00(.04)
My teacher gives us time to explain our ideas.	1.90(.02)	1.90(.02)	3.64(.04)	2.13(.03)	-0.27(.02)	-2.62(.03)
Students speak up and share their ideas about class work.	1.32(.02)	1.32(.02)	2.76(.03)	1.48(.02)	-0.42(.02)	-2.18(.02)
My teacher respects my ideas and suggestions.	1.97(.02)	1.97(.02)	3.60(.04)	2.21(.03)	-0.2(.02)	-2.35(.03)
My teacher takes the time to summarize what we learn each day.	1.45(.02)	1.45(.02)	2.61(.03)	1.25(.02)	-0.64(.02)	-2.42(.03)
My teacher checks to make sure we understand what s/he is teaching us.	2.24(.02)	2.24(.02)	4.50(.06)	2.94(.04)	0.44(.02)	-1.96(.03)
We get helpful comments to let us know what we did wrong on assignments.	1.65(.02)	1.65(.02)	3.18(.04)	1.9(.03)	-0.04(.02)	-2.08(.03)
The comments that I get on my work in this class help me understand how to improve.	1.68(.02)	1.68(.02)	3.35(.04)	2.04(.03)	-0.02(.02)	-2.05(.03)

Table 2.4: Tripod survey latent regression estimates

	Est.	SE
Fixed Effects		
Percentage Special education students	1.38	0.13
Class-average student-reported effort	1.73	0.05
Percentage minority students	0.16	0.01
Average prior year test score	0.13	0.02
Random Effects		
Class-level variance	0.20	0.01
Student-level variance	1.00	—

Table 2.5: Results for SGP simulated data example

Data-Generating Specifications						Results		
N_j	J	$r(\eta_p, \eta_c)$	$r(\vartheta_p, \vartheta_c)$	ICC_p	ICC_c	$r(M, C)$	$r(C, \hat{C})$	ρ_C
20	250	0.7	0.75	0.25	0.35	0.96	0.97	0.81
20	250	0.7	0.75	0.2	0.3	0.96	0.95	0.75
20	250	0.8	0.85	0.25	0.35	0.93	0.93	0.76
20	250	0.8	0.85	0.2	0.3	0.93	0.94	0.74
20	500	0.7	0.75	0.25	0.35	0.96	0.97	0.78
20	500	0.7	0.75	0.2	0.3	0.96	0.97	0.77
20	500	0.8	0.85	0.25	0.35	0.94	0.96	0.76
20	500	0.8	0.85	0.2	0.3	0.92	0.94	0.68
30	250	0.7	0.75	0.25	0.35	0.96	0.98	0.88
30	250	0.7	0.75	0.2	0.3	0.95	0.98	0.84
30	250	0.8	0.85	0.25	0.35	0.93	0.97	0.84
30	250	0.8	0.85	0.2	0.3	0.93	0.96	0.77
30	500	0.7	0.75	0.25	0.35	0.96	0.98	0.86
30	500	0.7	0.75	0.2	0.3	0.95	0.98	0.85
30	500	0.8	0.85	0.25	0.35	0.94	0.97	0.84
30	500	0.8	0.85	0.2	0.3	0.93	0.96	0.80

Figure 2.1: Smarter Balanced multilevel testlet item factor model

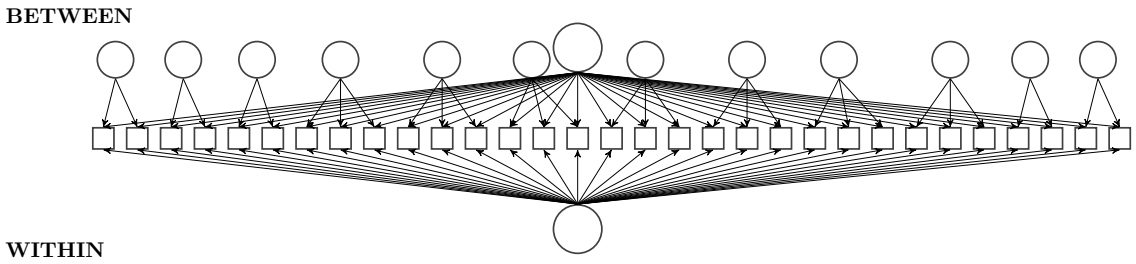
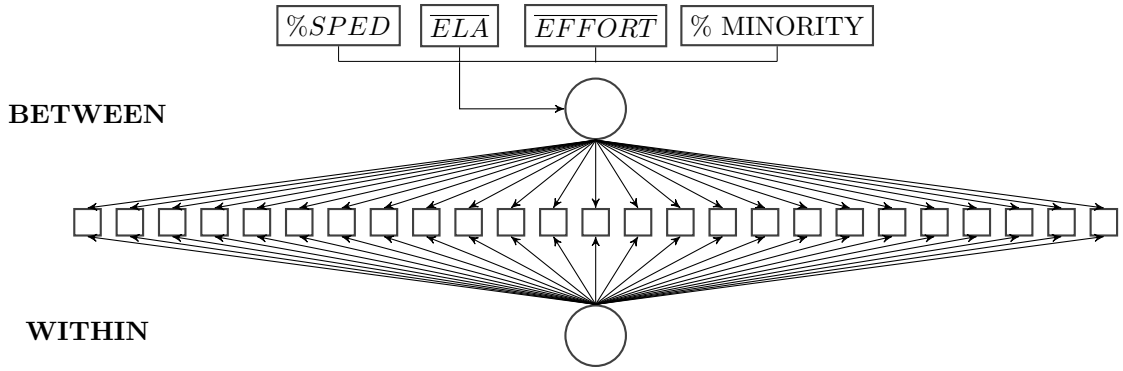
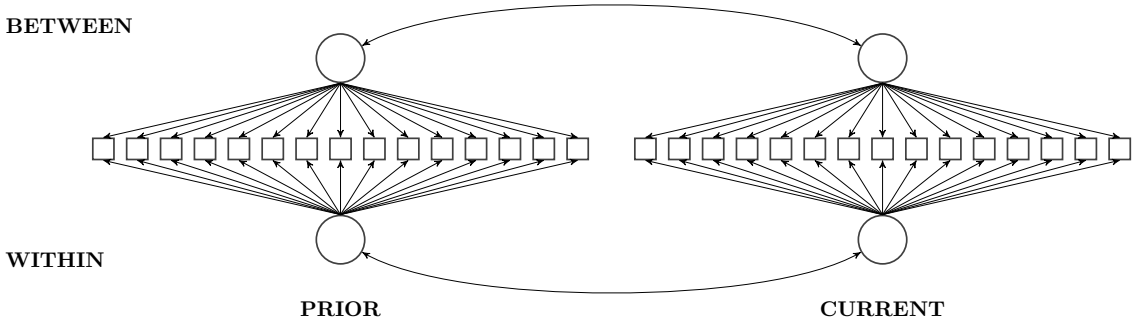


Figure 2.2: Path diagram for student surveys of teacher effectiveness



The classroom-level latent teacher practice dimension is regressed on the percentage of students in the class who are in a special education program, average prior year test score, class average of a student-reported measure of effort put forth in class, and the percentage of minority students in class.

Figure 2.3: Path diagram for cluster growth percentiles



CHAPTER 3

Limited-information goodness-of-fit testing of multilevel item factor analysis models

3.1 Abstract

Multilevel item factor analysis, an extension of widely-used item factor analysis models (IFA; Wirth & Edwards, 2007), has grown in popularity in the last ten years. Model fit assessment for multilevel IFA models, however, is still in its infancy. In this study, I first discuss the relationship between multilevel IFA models with balanced clustered data and single-level item bifactor models, and then I propose a limited-information goodness of fit statistic for multilevel IFA models. I demonstrate theoretically and with a simulated data example that the multilevel IFA model can be re-parameterized into a single-level item bifactor model that is fit to the group-level data. By re-specifying the multilevel item factor analysis model to be estimated in a single-level context, I am able to utilize the limited-information M_2 statistic proposed by Maydeu-Olivares and Joe (2006) for single-level factor models. Additionally, I propose a Reduced M_2 statistic to isolate the presence of item-level misfit. Through a series of simulation studies, I found that the M_2 statistic is sensitive to the examined misspecifications of the item model, but that the proposed Reduced M_2 is slightly conservative (with Type I error rates consistently below the nominal level).

3.2 Introduction

Traditional item factor analysis (IFA) models can be thought of as two-level models, as the analyzed item responses are nested within respondents. However, most IFA models do not consider a nested structure of the individuals into higher-level units.

Psychological and educational data frequently have such a nested data structure, such as data collected within classrooms and schools or data collected using multi-stage sampling procedures. Multilevel IFA models, which appropriately analyze data by taking into account both within- and between-cluster variations of the data, have been developed and refined over the last 15 years. Fox and Glas (2001) and Fox (2010) have proposed a multilevel item response theory model that can accommodate dichotomous and polytomous items, and is estimated using Markov chain Monte Carlo (MCMC) methods. Kamata (2001) described a hierarchical generalized linear model (HGLM) approach to multilevel item response theory model that permits investigation of the variation of students' performance across groups, and the interactive effect of person and group-characteristic variables. Additionally, frameworks uniting generalized linear mixed models, multilevel factor models, and item response models have been proposed by Rabe-Hesketh et al. (2004).

In this study, I discuss a flexible framework of multilevel IFA models that can be estimated using full-information maximum likelihood methods (Kuhfeld, Cai, & Monroe, in preparation). There are two primary aims of this study. I first establish that multilevel unidimensional IFA models with balanced cluster data can be re-parameterized as single-level item bifactor model that contains specific factors to account for the nesting of individuals in level-2 units. The relationship between these two models is established theoretically, and demonstrated with a small simulation study.

Secondly, I utilize the established relationship between models to examine the utility of the limited-information overall goodness-of-fit statistic M_2 (Maydeu-Olivares & Joe, 2006) to detect model misfit in multilevel item response data that have been reformatted into a dataset where groups are considered the level-1 unit. Various M_2-type statistics have been previously developed and applied to single-level item bifactor models (Cai & Hansen, 2013). I compare the existing M_2 statistic to a modified M_2 that is proposed in this paper. I believe that this work represents the first expansion of the M_2 statistic to multilevel item response data.

The rest of this study is organized as follows. I begin by demonstrating the multilevel IFA model with balanced clustered data can be re-parametrized as an item bifactor

model that contains specific factors for level-1 units. Next, I describe the development of M_2 in single-level IFA models, and propose a modified M_2 statistic for multilevel item response data with balanced clusters. Subsequently, I evaluate the performance of the two M_2 statistics through a series of simulation studies. Finally, I discuss some of the limitations of this research and opportunities to further develop this work.

3.3 Item factor analysis (IFA) models

3.3.1 Multilevel item factor analysis (MLIFA)

Let there be $p = 1, ..., n$ items. Let there be $i = 1, ..., N_j$ individuals in level-2 unit j, with $j = 1, ..., \mathrm{J}$ independent groups. The overall sample size is $\mathrm{N} = \sum_{j=1}^{\mathrm{J}} \mathrm{N}_j$. It can be assumed that y_{pij} takes integer values from $(0, \ldots, K_p - 1)$. Let the $n \times 1$ vector of item responses from respondent i in group j be $\mathbf{y}_{ij} = (y_{1ij}, \ldots, y_{pij}, \ldots, y_{nij})$. In this model, the latent variables for individual i in group j are partitioned into two mutually exclusive parts: $\boldsymbol{\theta}_{ij} = (\vartheta_j, \eta_{ij})$, where ϑ_j is the group-level (level-2) latent variable, and η_{ij} is the individual-level (level-1) latent variable.

The multilevel item factor analysis model specifies the conditional distribution of $\mathbf{y}_{ij}$ given the latent variables. In this study, I focus on items with two response categories, which can be coded as correct and incorrect. The 2-parameter logistic (2PL) model is used to specify the conditional probability of a correct response

$$P(y_{pij} = 1|\vartheta_j, \eta_{ij}) = \frac{1}{1 + \exp\left[-\left(c_p + a_p^B \vartheta_j + a_p \eta_{ij}\right)\right]}, \tag{3.1}$$

where y_{pij} is a Bernoulli random variable representing the response to item p from individual i in group j, c_p is the item intercept, and a_p and a_p^B are the level-1 and level-2 item slopes. The conditional probability of an incorrect response is $P(y_{pij} = 0|\vartheta_j, \eta_{ij}) = 1 - P(y_{pij} = 1|\vartheta_j, \eta_{ij})$.

The conditional distribution for an observed response y_{pij} is a multinomial distri-

bution with trial size 1 in K_p cells, and the conditional density is

$$f_{\boldsymbol{\gamma}}(y_{pij}|\vartheta_j, \eta_{ij}) = \prod_{k=0}^{K_p-1} P(y_{pij} = k|\vartheta_j, \eta_{ij})^{\chi_k(y_{pij})}, \tag{3.2}$$

where $\boldsymbol{\gamma}$ is the $d \times 1$ vector that collects together all free parameters in the model, and $\chi_k(y_{pij})$ is an indicator function that equals 1 if and only if $y_{pij} = k$ and 0 otherwise. Assuming conditional independence of item responses on level-1 latent variables, the conditional distribution of $\mathbf{y}_{ij}$ given ϑ_j and η_{ij} can be written as a product

$$f_{\boldsymbol{\gamma}}(\mathbf{y}_{ij}|\vartheta_j, \eta_{ij}) = \prod_{p=1}^{n} f_{\boldsymbol{\gamma}}(y_{pij}|\vartheta_j, \eta_{ij}). \tag{3.3}$$

The level-1 latent variables can be integrated out as

$$f_{\boldsymbol{\gamma}}(\mathbf{y}_{ij}|\vartheta_j) = \int \prod_{p=1}^{n} f_{\boldsymbol{\gamma}}(y_{pij}|\vartheta_j, \eta_{ij}) f(\eta_{ij}) d\eta_{ij}. \tag{3.4}$$

Assuming further conditional independence of level-1 units on ϑ_j, the marginal distribution of all level-2 unit is

$$f_{\boldsymbol{\gamma}}(\mathbf{y}_j) = \int \prod_{i=1}^{N_j} \left[\int \prod_{p=1}^{n} f_{\boldsymbol{\gamma}}(y_{pij}|\vartheta_j, \eta_{ij}) f(\eta_{ij}) d\eta_{ij} \right] f(\vartheta_j) d\vartheta_j. \tag{3.5}$$

Once the response pattern are observed (and considered fixed), the marginal likelihood of $\mathbf{y}_j$ is defined as

$$L(\boldsymbol{\gamma}|\mathbf{Y}_j) = f_{\boldsymbol{\gamma}}(\mathbf{Y}_j). \tag{3.6}$$

Invoking the assumption of independent level-2 units, the marginal log-likelihood is $\sum_{j=1}^{J} \log L(\boldsymbol{\gamma}|\mathbf{Y}_j)$. The value of the marginal likelihood can be approximated to arbitrary precision using numerical quadrature, i.e.,

$$L\left(\boldsymbol{\gamma}|\mathbf{Y}_j\right) \approx \sum_{q=1}^{Q} \prod_{i=1}^{N_j} \left[\sum_{q=1}^{Q} \prod_{p=1}^{n} f_{\boldsymbol{\gamma}}(y_{pij}|X_q, X_q) W_q \right] W_q, \tag{3.7}$$

where X_q and W_q are the quadrature node and weight, respectively. Numerically optimizing the marginal likelihood leads to the maximum likelihood estimate (MLE)

of $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$.

3.3.2 Single-level item bifactor analysis model

I now introduce the notation used for the item bifactor model. Let the response from individual i to item p be y_{pi}, where y_{pi} has K_p response categories, so that $y_{pij} \in (0, 1)$.

The bifactor model specifies the conditional probability for the response to item p with K_p categories from individual i. The 2-parameter logistic (2PL) model is used to specify the conditional probability of a correct response

$$P(y_{pi} = 1|\eta_i^g, \eta_i^s) = \frac{1}{1 + \exp\left[-\left(c_p + a_p\eta_i^g + a_p^s\eta_i^s\right)\right]}, \tag{3.8}$$

where the latent dimension η_i^g is the general dimension with slope a_p, η_i^s is the sth specific dimension with slope a_p^S, and c_p is the item intercept. Note that the item p is permitted to load on at most one specific dimension s.

The marginal likelihood of $\boldsymbol{\gamma}$ given the observed data $\mathbf{y}_i$ is

$$L\left(\boldsymbol{\gamma}|\mathbf{y}_i\right) = \int \prod_{s=1}^{S} \left[\int \prod_{j \in \mathcal{I}_s} f_{\boldsymbol{\gamma}}(y_{pi}|\eta_i^g, \eta_i^s) f(\eta_i^s) d\eta_i^s \right] f(\eta_i^g) d\eta_i^g. \tag{3.9}$$

As described by Cai, Yang, and Hansen (2011), the marginal likelihood shown in Equation 3.9 is a series of iterated integrals whose dimensionality is at most two.

Note that the bifactor marginal likelihood shown in Equation 3.9 is structured parallel to Equation 3.5. Within the first integral in the multilevel marginal likelihood, there is a product across all of the level-1 units. In Equation 3.9, there is a product of the specific factors integrals. It is clear that the level-1 units and specific factors play a similar role in the marginal likelihoods. Furthermore, the level-2 latent factor ϑ_j in the multilevel IFA model is parallel to the general factor η^g in the bifactor model. Assuming balanced clustered data, it is shown in the subsequent section that it is possible re-parameterize a multilevel IFA model as a single-level bifactor model where there is a specific factor for each of the level-1 units (e.g., N_j specific factors).

3.3.3 Refitting the multilevel item factor analysis model as a single-level bifactor model

A small simulation data example demonstrates that the multilevel unidimensional IFA with balanced clusters and single-level item bifactor model provide the same item parameter results when the appropriate set of constraints are in place. The multilevel IFA model is implemented in flexMIRT® item response modeling software version 3 (Cai, 2015), which is used in both data generation and model fitting.

The data generating model is a multilevel unidimensional IFA model and contains five items, all of which load both on the within-level and between-level latent factors. All items are dichotomous, and the item responses are simulated under a graded response model for two categories. The generating parameter values are shown in Table 3.1. The number of simulees within a level-2 unit (N_j) is 5, with $J = 200$ level-2 units. There are a total of $N = 1000$ simulated individuals. Data are organized in a $N \times n$ matrix, so that the $n = 5$ item responses for each individual are included as a row of data. A cluster ID variable is also included to specify which individuals are within the same level-2 unit.

First, a multilevel IFA model is fit to the multilevel generated item response data using the Bock-Aitken EM algorithm (Bock & Aitkin, 1981). The model took under five seconds to estimate. The estimated parameters are shown in Table 3.2. In this model, there are $d = 15$ free item parameters. The -2log-likelihood estimate for this model is 5397.04.

Secondly, the data are re-formatted so that the vector of all of the item responses with a level-2 unit are combined to be a single row of data. The re-formatted dataset has J rows and $n_{bf} = n \times N_j$ columns. In this re-organization, each individual in a level-2 unit is arbitrarily assigned to a placement (e.g., $i = 1, 2, ..., N_j$) within the level-2 unit. The reformatted data has $n_{bf} = 25$ pseudo-items, which represent the combined influence of the original pth item and the set of individuals assigned to the ith placement.

A constrained item bifactor model is estimated based on the re-formatted data.

The general factor η^g represents the level-2 influence, and the N_j specific factors $(\eta^{s_1}, ..., \eta^{s_{N_j}})$ represent the within-level influence. A set of equality constraints are put in place so that only $d = 15$ item parameters are estimated. The computation time, which included the estimation of the model as well as the calculation of the Jacobian and weight matrices, was a total of 45 seconds.

Table 3.3 displays the results of the constrained item bifactor model estimation with $n_{bf} = 25$ items. In this table, the items are labeled based on the "true" item (e.g., v1, v2) as well as which individual in the group is responding (e.g., p1, p2). The estimated parameters are equivalent to those in Table 3.2, and the -2log-likelihood estimate is also equivalent (5397.04).

Having established that I can estimate the multilevel IFA model as a single-level bifactor model with a set of constraints in place, I now turn to the examination of model goodness-of-fit.

3.4 Limited-information goodness-of-fit testing

3.4.1 Multivariate multinomial distributions

I first describe the development of M_2 statistic in the context of single-level IFA models. For n items, the item factor analysis model generates a total of $C = 2^n$ cross-classifications or possible item response patterns in the form of a contingency table. Based on a sample of N respondents, let the observed proportion associated with pattern $\mathbf{y} = (y_1, ..., y_n)'$ be denoted as $p_{\mathbf{y}}$. Let $\pi_{\mathbf{y}}(\boldsymbol{\gamma})$ denote the marginal likelihood. The sampling model for this contingency table is a multinomial distribution with C cells and N trials. The multinomial log-likelihood for the item parameters $\boldsymbol{\gamma}$ from a single-level IFA model is proportional to

$$\log L(\boldsymbol{\gamma}) \propto N \sum_{\mathbf{y}} p_{\mathbf{y}} \log \pi_{\mathbf{y}}(\boldsymbol{\gamma}), \tag{3.10}$$

where the summation is over all C response patterns. As described earlier, maximization of the log-likelihood leads to the maximum marginal likelihood estimator $\hat{\boldsymbol{\gamma}}$.

Once the maximum marginal likelihood estimator $\hat{\boldsymbol{\gamma}}$ has been estimated, the model generates model-implied probabilities for each response pattern $\hat{\pi}_{\mathbf{y}} = \pi_{\mathbf{y}}(\hat{\boldsymbol{\gamma}})$, which can be collected into a $C \times 1$ vector $\hat{\boldsymbol{\pi}}_{\mathbf{y}}$. Similarly, let a $C \times 1$ vector $\boldsymbol{\pi}_*$ contain the true (population) response pattern probabilities. Let there be a $C \times 1$ vector $\mathbf{p}$ containing all of the observed proportions. In a simple case with 3 items, there are $2^3 = 8$ item response patterns, and the response pattern probabilities and observed proportions are:

$$\boldsymbol{\pi}_* = \begin{pmatrix} \pi_{000} \\ \pi_{001} \\ \pi_{010} \\ \pi_{011} \\ \pi_{100} \\ \pi_{101} \\ \pi_{110} \\ \hat{\pi}_{111} \end{pmatrix}, \qquad \hat{\boldsymbol{\pi}} = \begin{pmatrix} \hat{\pi}_{000} \\ \hat{\pi}_{001} \\ \hat{\pi}_{010} \\ \hat{\pi}_{011} \\ \hat{\pi}_{100} \\ \hat{\pi}_{101} \\ \hat{\pi}_{110} \\ \hat{\pi}_{111} \end{pmatrix} = \begin{pmatrix} \pi_{000}(\hat{\boldsymbol{\gamma}}) \\ \pi_{001}(\hat{\boldsymbol{\gamma}}) \\ \pi_{010}(\hat{\boldsymbol{\gamma}}) \\ \pi_{011}(\hat{\boldsymbol{\gamma}}) \\ \pi_{100}(\hat{\boldsymbol{\gamma}}) \\ \pi_{101}(\hat{\boldsymbol{\gamma}}) \\ \pi_{110}(\hat{\boldsymbol{\gamma}}) \\ \hat{\pi}_{111}(\hat{\boldsymbol{\gamma}}) \end{pmatrix}, \qquad \mathbf{p} = \begin{pmatrix} p_{000} \\ p_{001} \\ p_{010} \\ p_{011} \\ p_{100} \\ p_{101} \\ p_{110} \\ p_{111} \end{pmatrix}. \tag{3.11}$$

Exactly correct model specification (e.g., perfect model fit) in the population can be stated as there exists $\boldsymbol{\gamma}_*$ such that $\boldsymbol{\pi}(\boldsymbol{\gamma}_*) = \boldsymbol{\pi}_*$, where $\boldsymbol{\gamma}_*$ is the vector of true parameters to be estimated.

Under correct model specification, the maximum likelihood estimator is consistent, asymptotically normal and asymptotically efficient (Bishop, Fienberg, & Holland, 1975). This can be summarized as follows:

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*) \xrightarrow{D} \mathcal{N}_d(\mathbf{0}, \mathcal{F}^{-1}), \tag{3.12}$$

where $\mathcal{F} = \boldsymbol{\Delta}_*^{'} \left[\text{diag}(\boldsymbol{\pi}_*)\right]^{-1} \boldsymbol{\Delta}_*$ is the $d \times d$ Fisher information matrix, with the Jacobian matrix $\boldsymbol{\Delta}_*$ defined as the $C \times d$ matrix of all first-order partial derivatives of the response pattern probabilities with respect to the parameters, evaluated at the true parameters $\boldsymbol{\gamma}_*$:

$$\boldsymbol{\Delta}_* = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\gamma}_*)}{\partial \boldsymbol{\gamma}^{'}}. \tag{3.13}$$

3.4.2 Distribution of residuals under maximum likelihood estimation

It can be shown (e.g., Bishop et al., 1975) that the asymptotic distribution of $(\mathbf{p} - \boldsymbol{\pi}_*)$ is C-variate normal:

$$\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}_*) \xrightarrow{D} \mathcal{N}_C(\mathbf{0}, \boldsymbol{\Xi}), \tag{3.14}$$

where $\boldsymbol{\Xi} = \text{diag}(\boldsymbol{\pi}_*) - \boldsymbol{\pi}_*\boldsymbol{\pi}_*'$ is the multinomial covariance matrix. Under maximum likelihood estimation, the cell residual vector $(\mathbf{p} - \hat{\boldsymbol{\pi}})$ is asymptotically C-variate normal:

$$\sqrt{N}(\mathbf{p} - \hat{\boldsymbol{\pi}}) \xrightarrow{D} \mathcal{N}_C(\mathbf{0}, \boldsymbol{\Gamma}), \tag{3.15}$$

where $\boldsymbol{\Gamma} = \boldsymbol{\Xi} - \boldsymbol{\Delta}_*\mathcal{F}^{-1}\boldsymbol{\Delta}_*'$.

The model also generates implied marginal probabilities. In general, these probabilities correspond to the n sets of univariate and $n(n-1)/2$ sets of bivariate margins that can be obtained from the full C-dimensional contingency table using a reduction operator matrix (see e.g., Maydeu-Olivares & Joe, 2006). An example of these marginal probabilities for $n = 3$ items is given by

$$\hat{\boldsymbol{\pi}}_2 = \begin{pmatrix} \dot{\pi}_1 \\ \dot{\pi}_2 \\ \dot{\pi}_3 \\ \ddot{\pi}_{21} \\ \ddot{\pi}_{31} \\ \ddot{\pi}_{32} \end{pmatrix} = \mathbf{L}\hat{\boldsymbol{\pi}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\pi}_{000} \\ \hat{\pi}_{001} \\ \hat{\pi}_{010} \\ \hat{\pi}_{011} \\ \hat{\pi}_{100} \\ \hat{\pi}_{101} \\ \hat{\pi}_{110} \\ \hat{\pi}_{111} \end{pmatrix}, \tag{3.16}$$

where $\mathbf{L}$ is an $r \times C$ fixed operator matrix of 0s and 1s that reduces the response pattern probabilities and proportions into marginal probabilities and proportions up to the second order. For dichotomously scored item responses, $r = n + n(n+1)/2$. The $r \times 1$ vector $\hat{\boldsymbol{\pi}}_2 = \mathbf{L}\hat{\boldsymbol{\pi}} = \mathbf{L}\boldsymbol{\pi}(\hat{\boldsymbol{\gamma}}) = \boldsymbol{\pi}_2(\hat{\boldsymbol{\gamma}})$ contains all first and second order model-implied marginal probabilities, evaluated at the maximum likelihood estimates. By analogy, $\mathbf{p}_2 = \mathbf{L}\mathbf{p}$ is the vector of first and second order observed marginal proportions.

A requirement on $\mathbf{L}$ is that it has full row rank, r. This implies that the marginal residual vector $(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2) = \mathbf{L}(\mathbf{p} - \hat{\boldsymbol{\pi}})$ is a full rank linear transformation of the multinomial cell residual vector $(\mathbf{p} - \hat{\boldsymbol{\pi}})$. Consequently, the marginal residual vector $(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)$ is also asymptotically normal:

$$\sqrt{N}(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2) = \sqrt{N}\mathbf{L}(\mathbf{p} - \hat{\boldsymbol{\pi}}) \xrightarrow{D} \mathcal{N}_r(\mathbf{0}, \boldsymbol{\Gamma}_2), \tag{3.17}$$

where $\boldsymbol{\Gamma}_2 = \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}' = \mathbf{L}\boldsymbol{\Xi}\mathbf{L}' - \mathbf{L}\boldsymbol{\Delta}_*\mathcal{F}^{-1}\boldsymbol{\Delta}_*'\mathbf{L}' = \boldsymbol{\Xi}_2 - \boldsymbol{\Delta}_{2*}\mathcal{F}^{-1}\boldsymbol{\Delta}_{2*}'$, where $\boldsymbol{\Xi}_2 = \mathbf{L}\boldsymbol{\Xi}\mathbf{L}'$, and $\boldsymbol{\Delta}_{2*} = \mathbf{L}\boldsymbol{\Delta}_*$ is the Jacobian of the marginal probabilities:

$$\boldsymbol{\Delta}_{2*} = \mathbf{L}\frac{\partial \boldsymbol{\pi}(\boldsymbol{\gamma}_*)}{\partial \boldsymbol{\gamma}'} = \frac{\partial \mathbf{L}\boldsymbol{\pi}(\boldsymbol{\gamma}_*)}{\partial \boldsymbol{\gamma}'} = \frac{\partial \boldsymbol{\pi}_2(\boldsymbol{\gamma}_*)}{\partial \boldsymbol{\gamma}'}. \tag{3.18}$$

Let $\hat{\boldsymbol{\Xi}} = \text{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}'$ denote the multinomial covariance matrix evaluated at $\hat{\boldsymbol{\gamma}}$, and let $\hat{\boldsymbol{\Xi}}_2 = \mathbf{L}\hat{\boldsymbol{\Xi}}\mathbf{L}'$. The marginal Jacobian evaluated at $\hat{\boldsymbol{\gamma}}$ is

$$\hat{\boldsymbol{\Delta}}_2 = \mathbf{L}\frac{\partial \boldsymbol{\pi}_2(\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}'}. \tag{3.19}$$

The model must be locally identified from the first and second order marginal probabilities. This local identification is achieved if $\hat{\boldsymbol{\Delta}}_2$ has full column rank, d.

3.4.3 $\mathbf{M}_2$ statistic

Limited-information overall fit statistics such as Maydeu-Olivares and Joe (2006)'s M_2 have gained prominence recently. As opposed to full-information statistics that use the full response pattern cross-classifications to examine the fit of the model, limited-information fit statistics use residuals based on lower order (e.g., first and second order) margins of the contingency table. A central advantage of limited-information statistics is that these lower order margins are far better filled when compared to the sparse full contingency table. Additionally, the limited-information goodness of fit tests have been found to be more powerful than full-information tests (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Cai & Hansen, 2013).

When $\hat{\mathbf{\Delta}}_2$ has full column rank, the statistic

$$M_2 = \sqrt{N}(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)' \hat{\mathbf{\Omega}} (\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2), \tag{3.20}$$

where $\hat{\mathbf{\Omega}} = \hat{\mathbf{\Xi}}_2^{-1} - \hat{\mathbf{\Xi}}_2^{-1}\hat{\mathbf{\Delta}}_2(\hat{\mathbf{\Delta}}_2^{'}\hat{\mathbf{\Xi}}_2^{-1}\hat{\mathbf{\Delta}}_2)^{-1}\hat{\mathbf{\Delta}}_2^{'}\hat{\mathbf{\Xi}}_2^{-1}$ is asymptotically chi-square distributed with $r - d$ degrees-of-freedom under the null hypothesis that the model fits exactly in the population (Browne, 1984).

When the model does not fit exactly in the population, the quadratic form in M_2 allows for the computation of a Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1993) index. Let $\hat{\mathcal{F}} = \sqrt{N}(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)' \hat{\mathbf{\Omega}} (\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)$ be the observed noncentrality. As per Browne and Cudeck (1993), an unbiased estimate of the population noncentrality is $F_* = \hat{\mathcal{F}} - \mathrm{df}/N$. The sample RMSEA based on M_2 is defined as a measure of the per degree-of-freedom noncentrality:

$$\mathrm{RMSEA} = \sqrt{\max\left(\frac{F_*}{\mathrm{df}}, 0\right)} \tag{3.21}$$

3.4.4 A reduced M_2 statistic

The derivation of M_2 described in the previous section has been well-examined for single-level hierarchical multidimensional IFA models, which includes bifactor and testlet models (Cai & Hansen, 2013). However, the single-level bifactor model specification for estimation of multilevel data is different in an important way: the n_{bf} items are really pseudo-items that represent the influence of the n observed items as well as sets of individuals that are assigned to each of the groupings within a level-2 unit. In the bifactor set-up, the first N_j pseudo-items represent the responses to the first "true" item from the entire set of individuals in the level-2 unit. Therefore, I would expect that the sets of N_j pseudo-items that represent a single "true" item to be related to each other more than the bifactor model currently explains.

Given that the goal in this study is to provide goodness-of-fit information for the multilevel IFA model with n items, I propose a M_2 statistic that is calculated using by collapsing the bifactor model first- and second-order margins to obtain a reduced

set of probabilities corresponding the "true" n items, rather than the larger set of n_{bf} items. From here on out, I refer to the M_2 statistic described in Equation 3.20 as the Full M_2, and the M_2 statistic proposed below as the Reduced M_2.

For a set of n "true" items, the number of linearly independent first-order marginal residuals is equal to $r1^* = \sum_{p=1}^{n}(K_p - 1)$. The number of linearly independent second-order marginal residuals is equal to $r2^* = \sum_{l=2}^{n}\sum_{m=1}^{l-1}(K_l - 1)(K_m - 1)$. Thus, taken together, $r^* = r1^* + r2^*$. I use a set of two operator matrices, $\dot{\mathbf{L}}_1^*$ and $\dot{\mathbf{L}}_2^*$, to reduce residual vector corresponding to the first- and second-order observed and expected probability vectors from the bifactor model to obtain the residuals for set of n items

$$\dot{\mathbf{e}}^* = \begin{pmatrix} \dot{\mathbf{e}}_1^* \\ \dot{\mathbf{e}}_2^* \end{pmatrix} = \begin{pmatrix} \dot{\mathbf{L}}_1^* \\ \dot{\mathbf{L}}_2^* \end{pmatrix} \mathbf{e}^* = \begin{pmatrix} \dot{\mathbf{L}}_1^* & \dot{\mathbf{L}}_2^* \end{pmatrix} \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \end{pmatrix} - \begin{pmatrix} \dot{\mathbf{L}}_1^* & \dot{\mathbf{L}}_2^* \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\pi}}_1 \\ \hat{\boldsymbol{\pi}}_2 \end{pmatrix}, \tag{3.22}$$

where $\mathbf{p}_1$ and $\hat{\boldsymbol{\pi}}_1$ are vectors of the first-order proportions and model-implied cell probabilities, and $\mathbf{p}_2$ and $\hat{\boldsymbol{\pi}}_2$ are vectors of the second-order proportions and model-implied cell probabilities, respectively. $\dot{\mathbf{L}}_1^*$ is a $n \times r1^*$ operator matrix that collapses the cell proportions and model-implied cell probabilities into $r1^*$ first-order marginal proportions, and $\dot{\mathbf{L}}_2^*$ is a $n \times r2^*$ operator matrix that collapses the cell proportions and model-implied cell probabilities into $r2^*$ second-order marginal proportions. It is also clear that $\dot{\mathbf{e}}^*$ is made up of the first- and second-order marginal residuals $\dot{\mathbf{e}}_1^*$ and $\dot{\mathbf{e}}_2^*$.

Let $\dot{\mathbf{L}}^*$ be the stacked vector of the two operator matrices $\dot{\mathbf{L}}^* = (\dot{\mathbf{L}}_1^*, \dot{\mathbf{L}}_2^*)'$. $\dot{\mathbf{L}}^*$ must have full row rank, r^*. Correspondingly, I can further reduce the Jacobian and weight matrix

$$\hat{\boldsymbol{\Delta}}_2^* = \dot{\mathbf{L}}^* \hat{\boldsymbol{\Delta}}_2, \quad \hat{\boldsymbol{\Xi}}_2^* = \dot{\mathbf{L}}^* \hat{\boldsymbol{\Xi}}_2 \dot{\mathbf{L}}^{*\prime}, \tag{3.23}$$

where $\hat{\boldsymbol{\Delta}}_2^*$ is $r^* \times d$ and $\hat{\boldsymbol{\Xi}}_2^*$ is a $r^* \times r^*$. The Reduced M_2 is the defined as

$$M_2^* = \sqrt{N}(\dot{\mathbf{e}^*})' \hat{\boldsymbol{\Omega}}^* (\dot{\mathbf{e}^*}), \tag{3.24}$$

where $\hat{\boldsymbol{\Omega}}^* = \hat{\boldsymbol{\Xi}}^{*-1}_2 - \hat{\boldsymbol{\Xi}}^{*-1}_2 \hat{\boldsymbol{\Delta}}_2^* (\hat{\boldsymbol{\Delta}}_2^{*\prime} \hat{\boldsymbol{\Xi}}^{*-1}_2 \hat{\boldsymbol{\Delta}}_2^*)^{-1} \hat{\boldsymbol{\Delta}}_2^{*\prime} \hat{\boldsymbol{\Xi}}^{*-1}_2$ is asymptotically chi-square distributed with $r^* - d$ degrees-of-freedom under the null hypothesis that the model fits

exactly in the population. It should be noted that the N in Equation 3.24 refers to the number of level-2 units, as the groups are the unit of analysis in this model.

3.5 Simulation studies

In order to evaluate the performance of the Full and Reduced M_2 in the context of multilevel IFA modeling, I conducted a series of simulation studies. First, I evaluated the calibration of the test statistic when the fitted model was correctly specified (i.e., matched the generating model; the null condition). I then examined the power of Full and Reduced M_2 to detect a variety of model misspecifications.

For all simulation conditions, the true test length was $n = 10$ items, and there were $N_j = 5$ individuals per level-2 unit, resulting in a bifactor model with $n_{bf} = 50$ items. Data were simulated under three level-2 sample conditions ($J = 200, 500, 1000$).

Figure 3.1 presents path diagrams for the four model structures: a multilevel unidimensional model (top left panel), a multilevel correlated factors model (top right panel), a multilevel bifactor model with specific factors at the within-level (bottom left panel), and a multilevel bifactor model with specific factors at the between-level (bottom right panel). In each replication, multilevel item response data are generated under each model structure, and then data are re-formatted to allow for group (single-level) analysis, and finally the single-level bifactor model (which corresponds to the multilevel unidimensional model) is fit to the data. The unidimensional data generating model is the null condition. The remaining three data generating models represent three different model misspecifications.

For each data generating condition, 100 datasets were generated in three level-2 sample sizes ($J = 200, 500, 1000$). All data generation and model estimation were conducted with the flexMIRT item response modeling software, version 3 (Cai, 2015). The Full and Reduced M_2 statistic was calculated using R software (R Core Team, 2012) based on the $\hat{\mathbf{\Delta}}_2$ and $\hat{\mathbf{\Xi}}_2$ matrices outputted in the *-dbg* file outputted from flexMIRT.

3.6 Results

3.6.1 Calibration of the test statistic (type I error)

Results for the Full M_2 statistic under the null and misspecified conditions are shown in Table 3.4. The null conditions are shown in the first section. The mean, variance, and empirical rejection rates obtained for the Full M_2 statistic across replications are close to what would be expected. Results for the Reduced M_2 statistic under the null are shown in the first section of Table 3.5. The Reduced M_2 has Type I error rates consistently below the nominal level across the sample sizes.

Figure 3.2 presents the quantile-quantile (QQ) plots comparing the observed and expected distributions of the Full and Reduced M_2 test statistic. For the most part, there is a good match between the distributions. Additionally, two-tailed Kolmogorov-Smirnov tests were used to evaluate the extent to which the observed distribution of the M_2 statistics differed from the expected chi-square reference distribution. At the $\alpha = 0.05$ level, the Kolmogorov-Smirnov test was not significant under any of the null conditions.

3.6.2 Power to detect misspecifications

Next, I examine the Full M_2 results obtained when data were generated with misspecifications. As shown in the bottom three sections of Table 3.4, this model was rejected in the majority of replications and at all but the lowest α level. Therefore, it appears that Full M_2 is sensitive to these types of model misspecification.

Table 3.5 displays the misspecified conditions for the Reduced M_2. For the correlated factors and bifactor model where the misspecification occurs at level-1, the empirical rejection rates for the Reduced M_2 are similar to the nominal α levels, indicating that the Reduced M_2 is not very sensitive to these types of misspecification. The power of the Reduced M_2 to detect misspecifications increases as sample sizes increases, but power remains lower than the Full M_2. The empirical rejection rates for the bifactor model with specific factors at the between-level are higher relative to the other two misspecified conditions, but not as high as the Full M_2.

3.7 Conclusions

In this paper, I demonstrate the application of limited-information fit statistics to multilevel item factor analysis models with balanced cluster data. It is demonstrated analytically and through a simulated data example that multilevel unidimensional item factor analysis models can be re-parameterized as a single-level item bifactor model with a specific factor for each level-1 unit in the data. I examine the standard M_2 statistic (which I refer to as Full M_2) for item bifactor models (e.g., Cai & Hansen, 2013), as well as proposed a Reduced M_2 that further collapsed the observed and model-implied response pattern probabilities. Through simulation studies I found that the Full M_2 is well calibrated, closely matching its reference distribution. The Full M_2 was found to be sensitive to all three types of misspecification. By contrast, I found that the Reduced M_2 was lightly conservative (with Type I error rates consistently below the nominal level) and had low power to detect misspecifications.

The current research is not without limitations. First, I have only focused on simulated data with balanced clusters (e.g., equal level-1 size for all groups). This choice was made because it allows for a very straightforward alignment between the multilevel IFA and single-level item bifactor model. However, this is unrealistic condition under real data collections, and therefore further work will be needed to generalize these findings to unbalanced data. Second, I have focused on the application of M_2 to dichotomous item response data. However, it would be beneficial to examine the performance of the test statistic with polytomous models. Third, the collapsing approach used in the Reduced M_2 will not work with all form lengths. For example, a multilevel IFA model with five items would have 15 freely estimated parameters and 15 first- and second-order margins. The degrees of freedom of the Reduced M_2 statistic would $df = 15 - 15 = 0$, and therefore it would not be possible to locally identify the Reduced M_2 from the first- and second-order margins.

Given that multilevel IFA models can now be efficiently estimated in commercially available item response modeling software (Cai, 2015), it is expected that these models will be more widely used over time, thus increasing the need to evaluate the fit of the models. This work represents a preliminary step into the assessment of goodness of fit

for multilevel item factor analysis, and I hope the initial evidence gathered here regarding limited-information goodness-of-fit testing for multilevel IFA models can prompt additional research.

Table 3.1: Data generating item parameters for the multilevel item factor analysis model (n=5)

Item	c	a^B	a^W
v1	2	0.5	1.0
v2	1	0.7	1.0
v3	0	0.8	1.0
v4	-1	0.6	1.0
v5	-2	0.75	1.0
	Mean	0(—)	0(—)
	Variance	1.0(—)	1.0(—)

Table 3.2: Estimated item parameters for the multilevel item factor analysis model (n=5)

Item	c	a^B	a^W
v1	1.85(0.12)	0.27(0.14)	0.65(0.19)
v2	1.07(0.11)	0.63(0.14)	1.00(0.22)
v3	0.02(0.08)	0.91(0.14)	1.05(0.22)
v4	-0.89(0.09)	0.67(0.12)	0.69(0.16)
v5	-2.01(0.17)	0.52(0.15)	1.05(0.25)
	Mean	0(—)	0(—)
	Variance	1.0(—)	1.0(—)

Table 3.3: Estimated item parameters for the item bifactor analysis model (n_{bf}=25)

Item	c	a_G	a_{s_1}	a_{s_2}	a_{s_3}	a_{s_4}	a_{s_5}
v1p1	1.85(.12)	0.27(.15)	0.65(.18)	0(—)	0(—)	0(—)	0(—)
v1p2	1.85(.12)	0.27(.15)	0(—)	0.65(.18)	0(—)	0(—)	0(—)
v1p3	1.85(.12)	0.27(.15)	0(—)	0(—)	0.65(.18)	0(—)	0(—)
v1p4	1.85(.12)	0.27(.15)	0(—)	0(—)	0(—)	0.65(.18)	0(—)
v1p5	1.85(.12)	0.27(.15)	0(—)	0(—)	0(—)	0(—)	0.65(.18)
v2p1	1.07(.12)	0.63(.13)	1.00(.22)	0(—)	0(—)	0(—)	0(—)
v2p2	1.07(.12)	0.63(.13)	0(—)	1.00(.22)	0(—)	0(—)	0(—)
v2p3	1.07(.12)	0.63(.13)	0(—)	0(—)	1.00(.22)	0(—)	0(—)
v2p4	1.07(.12)	0.63(.13)	0(—)	0(—)	0(—)	1.00(.22)	0(—)
v2p5	1.07(.12)	0.63(.13)	0(—)	0(—)	0(—)	0(—)	1.00(.22)
v3p1	0.02(.11)	0.91(.15)	1.05(.23)	0(—)	0(—)	0(—)	0(—)
v3p2	0.02(.11)	0.91(.15)	0(—)	1.05(.23)	0(—)	0(—)	0(—)
v3p3	0.02(.11)	0.91(.15)	0(—)	0(—)	1.05(.23)	0(—)	0(—)
v3p4	0.02(.11)	0.91(.15)	0(—)	0(—)	0(—)	1.05(.23)	0(—)
v3p5	0.02(.11)	0.91(.15)	0(—)	0(—)	0(—)	0(—)	1.05(.23)
v4p1	-0.89(.10)	0.67(.12)	0.69(.17)	0(—)	0(—)	0(—)	0(—)
v4p2	-0.89(.10)	0.67(.12)	0(—)	0.69(.17)	0(—)	0(—)	0(—)
v4p3	-0.89(.10)	0.67(.12)	0(—)	0(—)	0.69(.17)	0(—)	0(—)
v4p4	-0.89(.10)	0.67(.12)	0(—)	0(—)	0(—)	0.69(.17)	0(—)
v4p5	-0.89(.10)	0.67(.12)	0(—)	0(—)	0(—)	0(—)	0.69(.17)
v5p1	-2.01(.18)	0.52(.15)	1.05(.28)	0(—)	0(—)	0(—)	0(—)
v5p2	-2.01(.18)	0.52(.15)	0(—)	1.05(.28)	0(—)	0(—)	0(—)
v5p3	-2.01(.18)	0.52(.15)	0(—)	0(—)	1.05(.28)	0(—)	0(—)
v5p4	-2.01(.18)	0.52(.15)	0(—)	0(—)	0(—)	1.05(.28)	0(—)
v5p5	-2.01(.18)	0.52(.15)	0(—)	0(—)	0(—)	0(—)	1.05(.28)
	Mean	0(—)	0(—)	0(—)	0(—)	0(—)	0(—)
	Variance	1.0(—)	1.0(—)	1.0(—)	1.0(—)	1.0(—)	1.0(—)

Table 3.4: Simulation study results: Full M_2 calibration under null and misfit conditions

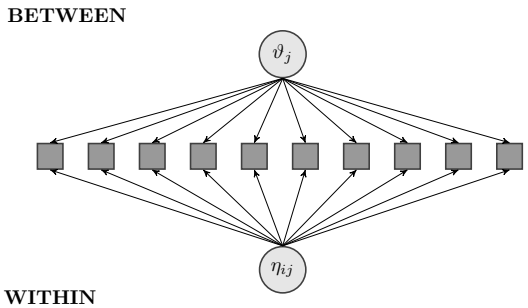
N	reps	df	M	V	Empirical Rejection Rate				
					0.2	0.15	0.1	0.05	0.01
Unidimensional (null) model									
200	100	1245	1238.89	2791.75	0.19	0.15	0.10	0.05	0.01
500	100	1245	1247.57	2509.51	0.25	0.15	0.06	0.03	0.01
1000	100	1245	1245.38	2185.91	0.18	0.11	0.08	0.06	0.00
Two correlated factors model									
200	100	1245	1403.55	4006.06	0.99	0.98	0.96	0.89	0.72
500	100	1245	1647.20	4945.48	1.00	1.00	1.00	1.00	1.00
1000	100	1245	2076.37	8950.71	1.00	1.00	1.00	1.00	1.00
Bifactor model (specific factors at level-1)									
200	100	1245	1836.98	9099.53	1.00	1.00	1.00	1.00	1.00
500	100	1245	2758.98	18873.83	1.00	1.00	1.00	1.00	1.00
1000	100	1245	4250.61	32302.33	1.00	1.00	1.00	1.00	1.00
Bifactor model (specific factors at level-2)									
200	100	1245	1338.68	2756.54	0.87	0.80	0.63	0.52	0.34
500	100	1245	1522.79	5436.61	1.00	1.00	1.00	1.00	1.00
1000	100	1245	1806.91	7669.73	1.00	1.00	1.00	1.00	1.00

Table 3.5: Simulation study results: Reduced M_2 calibration under null and misfit conditions

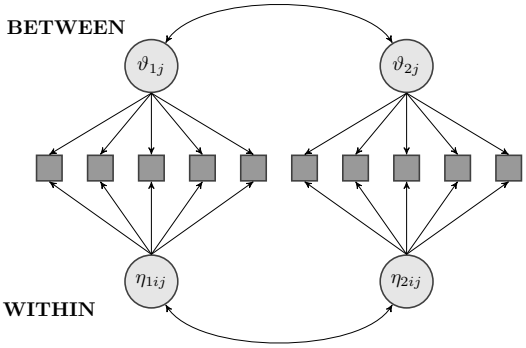
N	reps	df	M	V	Empirical Rejection Rate				
					0.2	0.15	0.1	0.05	0.01
Unidimensional (null) model									
200	100	25	24.56	37.72	0.17	0.13	0.07	0.02	0.00
500	100	25	24.59	48.47	0.19	0.16	0.08	0.03	0.01
1000	100	25	23.82	48.61	0.15	0.12	0.09	0.01	0.01
Two correlated factors model									
200	100	25	23.65	52.21	0.17	0.11	0.07	0.03	0.01
500	100	25	24.22	44.16	0.14	0.09	0.06	0.03	0.00
1000	100	25	28.76	63.44	0.34	0.30	0.21	0.16	0.03
Bifactor model (specific factors at level-1)									
200	100	25	21.26	40.81	0.11	0.06	0.06	0.00	0.00
500	100	25	23.01	45.93	0.09	0.09	0.07	0.04	0.00
1000	100	25	29.50	57.39	0.47	0.36	0.29	0.16	0.03
Bifactor model (specific factors at level-2)									
200	100	25	30.86	82.72	0.48	0.40	0.34	0.23	0.08
500	100	25	41.45	116.03	0.84	0.80	0.71	0.61	0.38
1000	100	25	59.00	233.84	0.98	0.98	0.97	0.95	0.78

Figure 3.1: Path diagrams for the four data generating models used in the simulation study

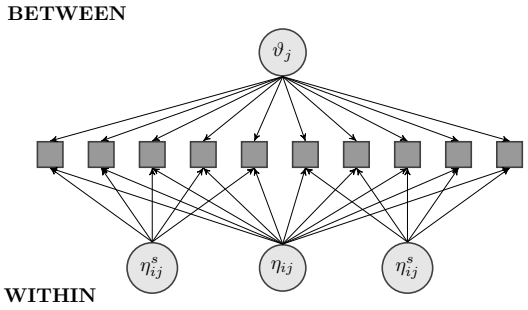
(a) Unidimensional model



(b) Correlated factors model



(c) Bifactor model (specific factors at level-1)



(d) Bifactor model (specific factors at level-2)

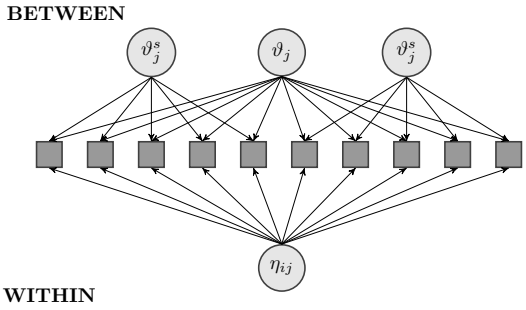
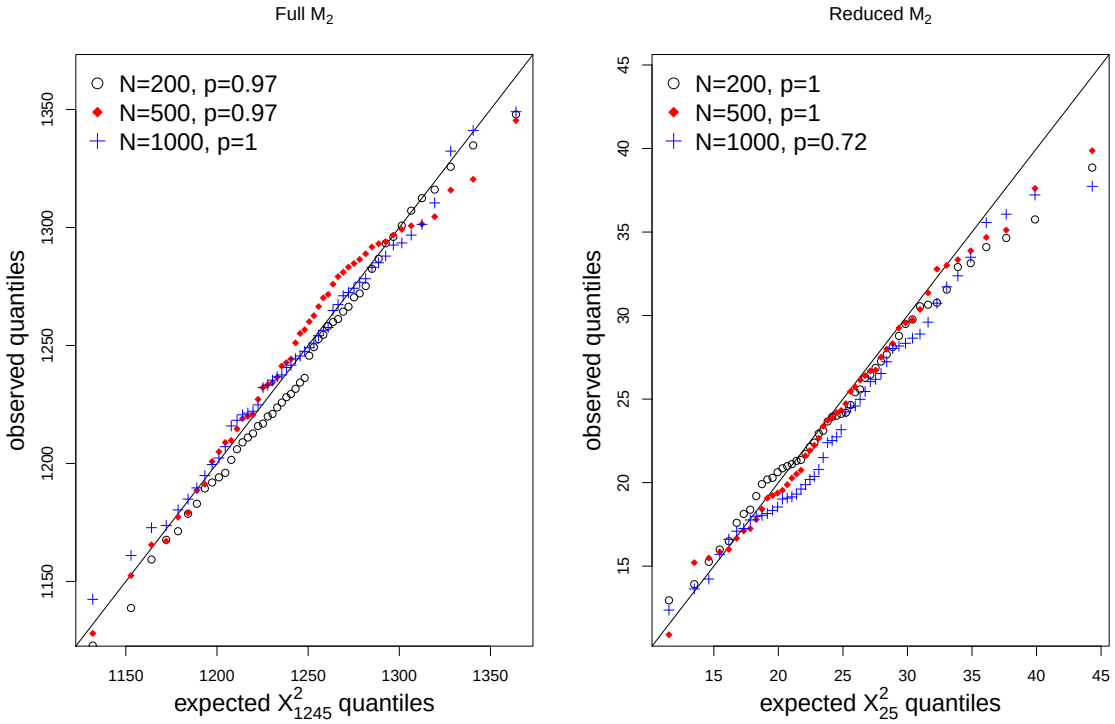


Figure 3.2: Simulation study results: Quantile-quantile plots of observed M_2 values and their reference chi-square distributions.



Open black circles indicate results for conditions with sample size N = 200, red solid diamonds indicate results for N = 500, and plus signs indicate N = 1000. Reported p-values are for a two-tailed Kolmogorov-Smirnov test of the equality of the observed χ^2 distributions with its corresponding reference distribution.

CHAPTER 4

A multilevel multidimensional plausible values approach for measuring teacher effectiveness

4.1 Abstract

This paper describes a multilevel multidimensional plausible values approach to account for measurement error while modeling the relationship between student perceptions of teacher practice and student academic growth. This model is illustrated in the context of predicting student learning in middle school English classrooms using seven latent teacher practices dimensions measured by the Tripod Survey. The multilevel multidimensional plausible values approach consists of two stages: (a) specifying and imputing sets of plausible values from a multilevel item bifactor measurement model implemented in flexMIRT® (Cai, 2015) and (b) fitting a multilevel model to predict student achievement by the imputed teacher practice values. In this paper, the multilevel measurement model is an item bifactor model containing a classroom-level overall teacher practice latent dimension as well as seven teacher practice specific factor dimensions. I compare this multilevel multidimensional plausible values approach with simpler modeling approaches to highlight the advantages and disadvantages of the plausible values approach.

4.2 Introduction

The increasing call for accountability has led to greater focus on teachers' instructional practices, which has been quantified using classroom observational protocols, student surveys, teacher logs, and teacher portfolios. However, the scores from these measures have generally not been found to explain a large proportion of classroom variation in

student learning gains (Kane et al., 2013; Kane & Cantrell, 2010). A possible explanation for the very modest associations between measures of teacher effectiveness and student achievement is the failure to account for measurement error in assessing teacher effectiveness. Using averages or summed scores to combine responses from an observational rubric or student survey can attenuate the correlation coefficients and regression estimates when relating scores to student learning. A promising approach involves a multilevel multidimensional plausible values method to appropriately account for measurement error while modeling the relationship between teacher practices and student academic growth. This paper focuses on one widely-used measure of teacher practice, the Tripod student perception survey, and compares the multilevel multidimensional plausible values method with simpler modeling approaches to examine the role of measurement error in the estimation of the association between multiple dimensions of teacher practice and student achievement.

Student surveys of instructional practice have recently gained prominence among researchers and policymakers as an inexpensive way to get feedback on what occurs inside the classroom. These surveys ask students about their opinions about specific teachers and specific classrooms. As of 2015, seven states mandate that teacher-level scores from student surveys be included in summative evaluation systems (Doherty & Jacobs, 2015). The Tripod survey is the most widely-used off-the-shelf student survey instrument (Kane & Cantrell, 2010). It was developed by Ron Ferguson at Harvard University, and is based upon classroom-level surveys developed by the Tripod Project for School Improvement (Ferguson, 2010). The Tripod student perceptions survey focuses primarily on what teachers do and how the classroom operates, which is operationalized as the Tripod 7Cs framework of teacher effectiveness. The seven dimensions of teacher practice measured by the Tripod survey are Care, Confer, Captivate, Clarify, Consolidate, Challenge, and Control.

The Tripod survey was included in the Bill and Melinda Gates Foundation's Measures of Effective Teaching (MET) study, a large study of students in thousands of classrooms in six urban districts in the United States. The MET project found that the Tripod survey was more reliable than student achievement gains or classroom ob-

servations (Kane & Staiger, 2012). Additionally, the MET project found that the overall Tripod index, computed as class-aggregated item mean scores, was correlated between .07 and .14 with value-added section scores, depending on whether the same section or different sections were used (Kane & Cantrell, 2010).

In this study, I go beyond answering the question of "does teacher effectiveness predict student achievement?" and identify which of the Tripod dimensions are most useful in predicting student learning. Specifically, I am interested in simultaneously modeling the relationships between the seven Tripod teacher practice dimensions and student academic achievement. This question appears on the face to be quite simple to answer, but there are important aspects of the data and assumptions of standard linear regression models that must be examined and addressed to properly relate latent teacher practice and student learning. The four central concerns that led to the choice of the multilevel multidimensional plausible values method modeling approach are described in the following section. Subsequently, I outline the methodological approach and describe how it addresses these four concerns.

4.3 Modeling considerations

There are four main considerations that drive the modeling framework used in this study: (a) measurement error, (b) collinearity among dimensions of teacher practice, (c) multilevel data structure, and (d) confounding variables.

Measurement error. Teaching practice cannot be measured directly, and it is well-known that all estimates of teacher practice (e.g., classroom observations, value-added scores, student surveys) will contain measurement error. Measurement error refers to the degree of imprecision or uncertainty in any assessment procedure. Standard linear regression models that do not account for uncertainty in the latent variables will produce estimates that are biased (Lu, Thomas, & Zumbo, 2005; Tucker, 1971). Latent variable models provide a very straight-forward framework for handling measurement error. Nonlinear multilevel latent variable models have been proposed to account for situations with multilevel structure where latent predictors are measured by categorical

manifest variables (see, for example Rabe-Hesketh et al., 2004). However, nonlinear multilevel latent variable models can be computational intensive when using full information maximum likelihood estimation, and require an advanced user to understand the model set-up. In this situation, it is tempting to use a latent framework such as item response theory (IRT) to produce latent score point estimates that can be plugged into a hierarchical linear model as a predictor. If teacher practice is measured without error, these scores can be plugged in directly to the outcomes model without adjustment. However, given that teacher practice is measured with some uncertainty, using point estimates will affect the precision and bias of estimated effects (Skrondal & Laake, 2001).

Correlations among predictors. Collinearity, which occurs when there are highly inter-correlated predictors, is a well-known issue in linear regression. As described by Raudenbush and Jean (2014), the Tripod dimensions are very highly correlated (ranging from 0.56 to 0.95), and so collinearity of the predictors presents a large problem when trying to address which of the seven Tripod dimensions of teacher practice is most predictive of student learning. One approach to dealing with the collinearity problem was described by Raudenbush and Jean (2014), who used a new method called the "multilevel variable selection model" to stabilize estimation and shrink unreliable coefficient estimates. Another approach would be to account for the high correlations among the 7Cs with a bifactor or higher-order measurement model that explicitly models the inter-correlations among dimensions through a general factor. The bifactor model assumes one common "teacher quality" factor underlies the variance of all of the observed items, and an additional set of orthogonal factors are specified to account for the unique sources of variance related to each of the 7Cs (Gibbons & Hedeker, 1992).

Multilevel structure. Hierarchical linear models (HLMs), also known as mixed models or random-effects models, have been used in social sciences to measure the relationship between constructs when data is collected in a nested structure, such as students nested within classrooms (Hox, 2002; Raudenbush & Bryk, 2002). In this study, the focus is on middle school students who are nested within English classrooms, who are in turn, nested within different schools.

Confounding variables. In addition to concerns about measurement error, I have reason to worry about generating estimates of teacher practice across a wide range of contexts, given the range of racial and socioeconomic composition across different classrooms. Students are not randomly assigned to classrooms, and therefore, there may be possible student and class-level confounding variables that relate to both teacher practice and student achievement. For example, the percentage of high-needs students in the classroom may affect both the teacher practices in the classroom, as well as affect the students' learning gains by the end of the year.

These four issues of measurement error, collinearity, multilevel data, and confounding variables have been addressed previously through the development of multilevel latent variable modeling frameworks, which have been proposed by Rabe-Hesketh et al. (2004) and Asparouhov and Muthén (2007), among others. However, when the observed variables are ordered-categorical Likert-type items, the estimation of the multidimensional multilevel latent variable models are typically extremely computationally demanding when the number of dimensions and the number of items is high. For this study, I am interested in simultaneously modeling the relationships between seven classroom-level dimensions of teacher practice and student academic achievement, where the seven dimensions are measured by a total of 36 Likert-type items. Given this data structure, widely used estimation approaches (limited-information weighted least squares, full-information maximum likelihood method with adaptive quadrature, or Markov Chain Monte Carlo methods) to fit a structural equation model with a seven-dimensional multilevel measurement model would be computationally difficult.

4.4 Multilevel multidimensional plausible values model

A promising alternative is to draw from the multiple imputation literature (Rubin, 1987, 1996) and the work of Mislevy, Beaton, Kaplan, and Sheehan (1992) and utilize multiple imputations of the latent trait as independent variables within a regression or hierarchical linear model. Mislevy et al. (1992) devised a two-stage imputation approach for use with the National Assessment of Educational Progress (NAEP), which is

followed by many large institutional surveys. These survey typically release individual proficiency data as a fixed number of multiple imputations. These imputations, also known as plausible values (PVs), are adjusted to account for measurement error in two ways. First, they are Monte Carlo draws from posterior proficiency distributions for each individual, and therefore incorporate measurement error and other sources of uncertainty. Second, the posterior distribution is conditioned on the individual responses to items on an assessment as well as a set of demographic and other background variables (Von Davier, Gonzalez, & Mislevy, 2009). This approach allows researchers to "borrow strength" from other individuals in the sample that are similar to a given individual by shrinking the latent estimate to a conditional mean based on the observed covariates. As described by Mislevy et al. (1992), it is important that the posterior distribution is conditioned on key covariates that are related to the latent trait of interest, as well as those covariates that will be included in the final model, including the outcome variable (e.g., student achievement).

Yang and Seltzer (2016) described a multilevel extension of this plausible values approach, which they called the "Multilevel Latent Variable Plausible Values" approach. The authors used a multilevel unidimensional measurement model in the first stage, and a fully Bayesian estimation approach is used to impute values from the first stage model. These plausible values are used as the predictor in a hierarchical linear outcomes model. The point estimates and standard errors in the final outcomes model are estimated employing the well-known formulas developed by Rubin (1987) to combine the multiple imputations.

In this study, I expand upon the "Multilevel Latent Variable Plausible Values" approach to examine the relationship between the seven theoretical dimensions of teacher practice (as measured by the Tripod student perception survey) and student academic achievement. This expansion consists of changing the unidimensional multiple imputation measurement model into a multidimensional multilevel item bifactor analysis model. The bifactor model is an approach to handle the situation where a general construct of interest (teacher effectiveness) is hypothesized to include several highly-related domains (e.g., caring, support, academic press). I chose to use a bifactor model in this

context rather than a unidimensional teacher practice dimension due to the desire to concurrently examine the relationship between multiple types of teacher practice and student achievement.

The bifactor model produces uncorrelated general and domain-specific group-level factors that can be examined as unique predictors of student achievement, even when all factors are in the same model (Chen, West, & Sousa, 2006). The bifactor model has been previously used with a different measure of teaching effectiveness, the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008), to conceptualize the association between domain-general and domain-specific aspects of teaching and preschool children's development (Hamre, Hatfield, Pianta, & Jamil, 2014). The multilevel extension of the bifactor model allows for the estimation of classroom "shared" perception of overall teacher practice, classroom "shared" perception of each of the 7Cs, as well as students' latent deviation from the class overall mean.

The multilevel multidimensional plausible values approach accounts for the four central concerns described in the previous section. The use of plausible values rather than point estimates to represent the latent independent variables incorporates measurement error in the outcome model. The bifactor structure accounts for the high inter-correlations among the Tripod dimensions by specifying that the overall dimension of teacher practice and the specific 7Cs dimensions are mutually orthogonal. The multilevel structure of the data is addressed by using multilevel models in both the measurement and outcomes model. Lastly, possible confounders are accounted for through conditioning on key covariates in the imputation model through a latent regression specification, as well as including those same covariates as predictors in the outcomes model.

However, this model is highly complex, and requires software that allows for the estimation of multilevel multidimensional item factor analysis model with observed categorical indicators, as well as the regression of the latent variables on observed covariates. Given this complexity, it is important to examine the degree to which this approach leads to different conclusions than other more widely-used scoring approaches. The standard approach to produce classroom-level teacher practice scores is to average

student item responses to estimate construct scale scores, then to average theses scores to obtain classroom means. Another alternative would be to rely on conventional IRT scale scores (e.g., Expected A Posteriori (EAP) scores) to get point estimates for the key latent predictors. A third alternative would be to estimate conditional EAP scores for the teacher practice predictor variables, where key classroom background covariates are included in the scoring model so that all scores are not shrunk to the same grand mean.

The overarching goal of the study is to demonstrate the usefulness of the multilevel multidimensional plausible value approach using an illustrative example of measuring the relationship between student perceptions of teacher practices and standardized test score in English/Language Arts (ELA). Using the Tripod survey collected as a part of the Measures of Effective teaching (MET) study, the following research questions were addressed:

1. **Research Question 1:** How similar are scores from the different scoring approaches? How much uncertainty do we see in the estimates of teacher practice?

2. **Research Question 2:** How do the resulting inferences about policy-relevant dimensions of teacher practices differ using standard scoring approaches from those based on a multilevel multidimensional plausible values approach?

In the following sections, I describe the design of the MET study and the measures collected. Second, I outline the multilevel multidimensional plausible values approach and the implementation of this approach using flexMIRT® (Cai, 2015) and the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015). Third, I present the results of the study. Finally, I conclude with a discussion of the limitations and challenges of the work and implications for policy and research.

4.5 Data and methods

4.5.1 Sample

The Measures of Effective Teaching (MET) study was conducted during two school years (2009-2010 and 2010-2011 years) in six districts in the United States. The purpose of this study was to examine the reliability and validity of the various measures of teaching effectiveness. A sample of approximately 3,000 teacher volunteers was recruited from six urban districts: New York City Department of Education, Charlotte-Mecklenburg Schools, Denver Public Schools, Memphis City Schools, Dallas Independent School District, and Hillsborough County Public Schools.

I focus on students in Grades 6 to 8 in English classrooms during the 2009-2010 school year. Students were included in this study if they were in a participating middle school English classroom that had at least five students who filled out the Tripod survey. The final analytical sample includes 13,989 students, within 884 classrooms taught by 463 teachers. Table 4.1 describes the sample at the student-, class-, and teacher-level. Thirteen percent of the students are classified as Gifted, 5% of students receive special education services, 13% are English learners (EL), and approximately 62% of students were eligible for free or reduced-price lunch (FRPL). Moreover, approximately 23% of students are White, approximately 26% are Black, approximately 41% are Hispanic, and 8% are Asian. Over half of the teachers in this sample have been teaching for seven or more years, and 29% have a master's degree or higher.

4.5.2 Measures

I use the secondary Tripod survey as a measure of teacher effectiveness (Ferguson, 2012). This measure was designed for middle and high school students, and asks students to rate many aspects of the teacher's behavior towards students. The Tripod items are organized under seven constructs, called the 7Cs: Care, Control, Clarify, Challenge, Captivate, Confer, and Consolidate. There are a total of 36 items in the secondary survey, and students are asked to rate teacher practices using a Likert-type response options with a 5-point scale (Totally Untrue; Mostly Untrue; Somewhat;

Mostly True; Totally True). Table 4.2 provides the wording of the items, as well as the 7s dimension that each item measures and the mean and standard deviation of item responses.

Tripod surveys were administered both on paper and online, with the participating schools choosing the mode of administration. Student surveys were collected during the 2009-10 school year prior to the state standardized test scores.

Student test scores on state-mandated exams were collected from administrative records for the first two years of the study (i.e., 2009–2010 and 2010–2011) and up to three years prior. The standardized test administered in 2010 (at the end of the 2009–10 school year) is the outcome variable in this model. Given that students in each school district took a separate state standardized exam, scores were standardized to have a mean of zero and a standard deviation of one within each grade and school district.

4.5.3 Multilevel multidimensional plausible values model specification

I now discuss the logic of the two-stage approach to measure the relationship between teacher practice and the end-of-year English/Language art standardized test scores. I term the first-stage model a multilevel measurement-imputation model, and the second-stage model a multilevel outcome model. I begin by first focusing on the specification of the multilevel measurement-imputation model, and then by describing the outcome model. The measurement model is depicted graphically in Figure 4.1.

Multilevel measurement-imputation model. Let there be $p = 1, \ldots, n$ items, and $i = 1, \ldots, N_j$ students in classroom j, with $j = 1, \ldots, J$ groups. Let the response from individual i in classroom j to item p be y_{pij}, where y_{pij} has K_p response categories, so that $y_{pij} \in (0, \ldots, K_p - 1)$. The overall sample size is $N = \sum_{j=1}^{J} N_j$.

In this model, the latent variables for individual i in classroom j are partitioned into two mutually exclusive parts: $\boldsymbol{\theta}_{ij} = (\boldsymbol{\vartheta}_j, \eta_{ij})$, where $\boldsymbol{\vartheta}_j = (\vartheta_j, \vartheta_{1j}, \vartheta_{2j}, \vartheta_{3j}, \vartheta_{4j}, \vartheta_{5j}, \vartheta_{6j}, \vartheta_{7j})$ is the vector of group-level (level-2) latent variables, and η_{ij} is the individual-level (level-1) latent variable. The multilevel item factor analysis model specifies the conditional probability for the response to item p with K_p categories from student i in classroom

j. I use a multidimensional extension of the graded response model (Samejima, 1969). Let the cumulative category response probabilities be

$$\begin{aligned} P(y_{pij} \geq 1|\boldsymbol{\vartheta}_j, \eta_{ij}) &= \frac{1}{1 + \exp\left[-\left(c_{p,k} + \mathbf{a}_p^B \boldsymbol{\vartheta}_j + a_p \eta_{ij}\right)\right]} \\ &\vdots \\ P(y_{pij} \geq K - 1|\boldsymbol{\vartheta}_j, \eta_{ij}) &= \frac{1}{1 + \exp\left[-\left(c_{p,K-1} + \mathbf{a}_p^B \boldsymbol{\vartheta}_j + a_p \eta_{ij}\right)\right]} \end{aligned} \tag{4.1}$$

The item parameters for item p include: a set of K-1 (strictly ordered) intercepts $c_{p,1}, \ldots, c_{p,K-1}$, the level-1 slope a_p, and a conformable vector of level-2 item slopes $\mathbf{a}_p^{B'}$. The slopes (or discrimination) parameters are analogous to item factor loadings. The category response probability is the difference between two adjacent cumulative probabilities

$$P(y_{pij} = k|\boldsymbol{\vartheta}_j, \eta_{ij}) = P(y_{pij} \geq k|\boldsymbol{\vartheta}_j, \eta_{ij}) - P(y_{pij} \geq k + 1|\boldsymbol{\vartheta}_j, \eta_{ij}) \tag{4.2}$$

where $P(y_{pij} \geq 0|\boldsymbol{\vartheta}_j, \eta_{ij})$ is equal to 1 and $P(y_{pij} \geq K_p|\boldsymbol{\vartheta}_j, \eta_{ij})$ is zero. In the multilevel bifactor model, an item always loads on the student-level general factor, the classroom-level general factor, and a single classroom-level specific factor. There are seven specific factors in this model at the classroom-level, which represent the seven Tripod dimensions of teacher practice. The general factors and the specific dimensions are jointly normally distributed and mutually orthogonal. Additional equality constraints are placed on the specific factor between-level item slopes so that each of the 7Cs specific factors can be seen as random deviations from the overall teacher effectiveness factor ϑ_j.

In the imputation model, it is necessary to condition on all of the variables included in the outcomes analysis model, including the outcome variable (Yang & Seltzer, 2016). In level-1 (e.g., within-classroom) model, I model η_{ij} as a function of the group mean

ϑ_j and a set of student background covariates

$$\eta_{ij} = \vartheta_j + \beta_{(M)1}\text{ELA09}_{ij} + \beta_{(M)2}\text{ELA10}_{ij} + \beta_{(M)3}\text{WHITE}_{ij} + \\ \beta_{(M)4}\text{FRPL}_{ij} + \beta_{(M)5}\text{GIFTED}_{ij} + \epsilon_{(M)ij}, \epsilon_{(M)ij} \sim N(0, \sigma^2_{(M)}) \tag{4.3}$$

In this model, ELA09_{ij} is the estimate of student prior end-of-year standardized English/Language Arts test score, and ELA10_{ij} is an estimate of student end-of-year standardized English/Language Arts test score (e.g., the outcome variable). The student end-of-year test scores are standardized within district, subject, and grade. Additionally, three student background characteristics are included: WHITE_{ij} is an indicator for whether the student is white, FRPL_{ij} is an indicator variable representing whether the student receives free or reduced price lunch (FPRL), and GIFTED_{ij} is an indicator for whether the student is labeled as gifted. This set of covariates was included because they are hypothesized to be related both to student ratings of teacher practice, as well as the outcome variable in the final model (end-of-year standardized English/Language Arts test score). Descriptive statistics for the student and classroom covariates are reported in Table 4.1.

All five covariates in Equation 4.3 are centered around their classroom means. This group-mean centering allows ϑ_j in Equation 4.3 to have a useful interpretation: ϑ_j is the mean rating of teacher practice in classroom j. This allows for investigation of the within-classroom and between-classroom relations between perceptions of teacher practice and student academic achievement. The student residual $\epsilon_{(M)ij}$ is assumed to be normally distributed with zero mean and variance $\sigma^2_{(M)}$.

The classroom overall teacher effectiveness latent variable ϑ_j is regressed on the class averages of the student covariates and the class average of the outcome variable ($\overline{\text{ELA10}}_j$):

$$\vartheta_j = \gamma_{(M)00} + \gamma_{(M)01}\overline{\text{ELA09}}_j + \gamma_{(M)02}\overline{\text{ELA10}}_j + \gamma_{(M)03}\overline{\text{SPED}}_j + \\ \gamma_{(M)04}\overline{\text{WHITE}}_j + \gamma_{(M)05}\overline{\text{FRPL}}_j + \gamma_{(M)06}\overline{\text{EL}}_j + \\ + \gamma_{(M)07}\overline{\text{GIFTED}}_j + \mu_{(M)j}, \quad \mu_{(M)j} \sim N(0, \tau_{(M)}) \tag{4.4}$$

The contextual variables that were included in this model are: class average prior year achievement ($\overline{\text{ELA09}}_j$), percentage students receiving special education services ($\overline{\text{SPED}}_j$), percentage of white students ($\overline{\text{WHITE}}_j$), percentage of students receiving free or reduced price lunch ($\overline{\text{FRPL}}_j$), percentage of English Learners ($\overline{\text{EL}}_j$), and percentage of gifted students ($\overline{\text{GIFTED}}_j$). All of the predictors in Equation 4.4 are grand-mean centered. The classroom-level residuals $\mu_{(M)j}$ is assumed to be normally distributed with means equal to zero and variance equal to $\tau_{(M)}$.

Scores are produced from the multilevel measurement model based on the posterior distribution of $\boldsymbol{\vartheta}$, conditional on the observed item responses $\mathbf{Y}$, and the observed covariates contained in the matrix $\mathbf{X}$, and the vector of all estimable item and/or structural parameters contained in $\boldsymbol{\gamma}$:

$$\text{P}(\boldsymbol{\vartheta}|\mathbf{Y},\mathbf{X},\boldsymbol{\gamma}) = \frac{f\left(\mathbf{Y}|\boldsymbol{\vartheta},\boldsymbol{\gamma}\right) f(\boldsymbol{\vartheta}|\mathbf{X})}{\int f\left(\mathbf{Y}|\boldsymbol{\vartheta},\boldsymbol{\gamma}\right) f(\boldsymbol{\vartheta}|\mathbf{X}) d\boldsymbol{\vartheta}}, \tag{4.5}$$

where $f\left(\mathbf{Y}|\boldsymbol{\vartheta},\boldsymbol{\gamma}\right)$ is the likelihood distribution of the item responses, and $f(\boldsymbol{\vartheta}|\mathbf{X})$ is the prior distribution. Plausible values are estimated as Monte Carlo draws from the posterior distribution in Equation 4.5, and the expected a posteriori (EAP) scores are estimated as the expectation of the posterior (Bock & Mislevy, 1982).

Multilevel outcome model. A hierarchical linear model (HLM) is specified to measure the relationship between end-of-year student achievement in English (ELA10_{ij}) and student covariates, classroom-level covariates, and the latent teacher practice estimates. The level-1 (within-classroom) outcome model is

$$\begin{aligned}\text{ELA10}_{ij} = {} & \beta_{(Y)0j} + \beta_{(Y)1j}\text{ELA09}_{ij} + \beta_{(Y)2j}\text{WHITE}_{ij} + \beta_{(Y)3j}\text{FRPL}_{ij} \\ & + \beta_{(Y)4j}\text{GIFTED}_{ij} + \epsilon_{(Y)ij}, \quad \epsilon_{(Y)ij} \sim N(0, \sigma^2_{(Y)})\end{aligned} \tag{4.6}$$

I distinguish the regression parameters in the hierarchical outcome model ($\beta_{(Y)j}$) from those in the level-2 imputation model ($\beta_{(M)}$) through the M and Y subscripts. The five student-level covariates are the same as in the student-level measurement latent regression model shown in Equation 4.3. In the within-classroom outcome model, all of the predictors are centered around group means so that $\beta_{(Y)0j}$ represents the

expected end-of-year standardized English/Language Arts test score for classroom j. The coefficients for the level-1 predictors are treated as fixed in my analyses.

At level-2, I model classroom-mean spring reading achievement ($\beta_{(Y)0j}$):

$$\begin{aligned}\beta_{(Y)0j} =& \gamma_{(Y)00} + \gamma_{(Y)01}\overline{\text{ELA09}}_j + \gamma_{(Y)02}\overline{\text{SPED}}_j + \\ &+ \gamma_{(Y)03}\overline{\text{WHITE}}_j + \gamma_{(Y)04}\overline{\text{FRPL}}_j + \gamma_{(Y)05}\overline{\text{EL}}_j + \gamma_{(Y)06}\overline{\text{GIFTED}}_j \\ &+ \gamma_{(Y)07}\text{OVERALL}_j + \gamma_{(Y)08}\text{CARE}_j + \gamma_{(Y)09}\text{CONTROL}_j \\ &+ \gamma_{(Y)10}\text{CLARIFY}_j + \gamma_{(Y)11}\text{CHALLENGE}_j + \gamma_{(Y)12}\text{CAPTIVATE}_j \\ &+ \gamma_{(Y)13}\text{CONFER}_j + \gamma_{(Y)14}\text{CONSOLIDATE}_j + \mu_{(Y)j},\ \mu_{(Y)j} \sim N(0, \tau_{(Y)})\end{aligned} \tag{4.7}$$

At level 2, $\beta_{(Y)0j}$ is modeled as a function of the grand mean, $\gamma_{(Y)00}$, class-level covariates, the estimates of teacher practice, and the random effect around the means, $\mu_{(Y)j}$. The random effect is assumed to be normally distributed with mean 0 and variance $\tau_{(Y)}$. In Equation 4.7, I am representing the estimates of latent teacher practice ($\vartheta_j, \vartheta_{1j}, \vartheta_{2j}, \vartheta_{3j}, \vartheta_{4j}, \vartheta_{5j}, \vartheta_{6j}, \vartheta_{7j}$) by their Tripod names (OVERALL$_j$, CARE$_j$, CONTROL$_j$, CLARIFY$_j$, CHALLENGE$_j$, CAPTIVATE$_j$, CONFER$_j$, CONSOLIDATE$_j$). The key parameters of interest in the outcome model are $\gamma_{(Y)07}$ through $\gamma_{(Y)14}$, which capture how differences between the J teachers with respect to the 7Cs relates to differences in class-mean end of year achievement, holding constant the predictors in the model. All of the variables in Equation 4.7 were centered around the grand means. Unobserved differences between the school districts were controlled for by including district fixed effects in the model.

4.5.4 Implementation

Four approaches were used to produce classroom-level teacher practice scores: (a) standardized class mean scores, (b) standard expected a posteriori (EAP) scores, (c) EAP scores that were from the multilevel item bifactor model with latent regression on classroom and teacher covariates, and (d) multilevel plausible values.

In the first scoring approach, student item responses are averaged to estimated construct scale scores, then these scores are aggregated to classroom means. Scores

are produced for all 7Cs, but an overall teacher effectiveness dimension is not included in the outcomes model given that the overall score will be highly correlated with each of the other 7Cs in this approach. The class mean scores are standardized to have a mean of zero and standard deviation of 1 within the sample.

The remaining scoring approaches rely on the specification of the multilevel measurement model. The multilevel item bifactor analysis measurement model with latent regression was estimated in flexMIRT® item response modeling software, version 3 (Cai, 2015) using the Metropolis-Hastings Robbins-Monro algorithm (Cai, 2010b). The item slopes for the general factor are fixed equal across levels, representing cross-level measurement invariance. This constraint allows for the between-level general latent variable (ϑ_j) to be interpreted as classroom "shared" perceptions of teacher effectiveness, while the within-level latent variable (η_{ij}) can be interpreted as the deviation from the classroom mean. Additionally, equality constraints are placed on the between-level specific factors, so that a single slope coefficient is estimated for each specific factor. The level-2 (between) variance for the general dimension was freely estimated, relative to a fixed within factor variance of 1.0. All of the specific factor variances are set to 1.0, and all of the factors in this model are uncorrelated. The latent regression equations shown in Equations 4.3 and 4.4 were estimated simultaneously with the item parameters in the flexMIRT calibration.

Item slope and factor mean/variance estimates for the multilevel item bifactor analysis model are presented in Table 4.3. The between-level general variance is estimated to be 0.30, implying an intraclass correlation (ICC) estimate of 0.23. The latent regression parameters are reported in Table 4.4. Students in classrooms with high percentages of special education students and English Learners reported significantly higher levels of teacher effectiveness. Additionally, the regression of the overall teacher effectiveness latent variable on the class-aggregated student test score variable ($\overline{\text{ELA10}}_j$) was significant, implying that a relationship in the outcomes model may be expected as well.

The second and third scoring approaches use classroom level Expected A Posteriori (EAP) scores, which are estimated using item parameters, regression coefficients,

and variance components fixed to the maximum likelihood estimates from flexMIRT calibration. For the jth classroom, let the EAP estimates be EAP(ϑ_j) and the corresponding standard errors be SE(ϑ_j), where EAP(ϑ_j) is the expectation of the posterior distribution of ϑ_j given in Equation 4.5. In the second scoring approach, EAP scores are produced with a common prior distribution, and so the scores for classroom are shrunk toward the grand mean teacher effectiveness value for the entire sample. In the third approach, which I refer to as EAP scoring with conditional means, the prior distribution $f(\boldsymbol{\vartheta}|\mathbf{X})$ is conditional on the observed classroom-level covariates $\mathbf{X}$. That is to say, in the EAP with conditional means scoring approach, EAP(ϑ_j) scores for classroom j are shrunk toward the expected value for classrooms who are similar in key ways (e.g., classes with similar student compositional characteristics). For the first three scoring approaches, scores were imported into R, and the multilevel outcome model specified in Equations 4.6 and 4.7 was estimated using lme4 package in R (Bates et al., 2015).

In the final scoring approach, I impute M=10 sets of values from the posterior in Equation 4.5. Each set of the imputed values, along with the observed student and classroom-level covariates, are termed an augmented data set. The multilevel outcome model was estimated using lme4 for each augmented data set, resulting in 10 estimates of the outcome parameters and 10 sets of parameter error variances. Rubin's (1987) formulas were used to combine the results across imputations and obtain an average estimate and error variance of the average for the parameters in the outcomes model. Letting $\gamma^m_{(Y)}$ be the estimate for a fixed effect of interest based on the mth set of imputed values, the average over the M=10 imputations is

$$\overline{\gamma}_{(Y)} = \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} \gamma^m_{(Y)}, \tag{4.8}$$

which gives the marginal estimate of $\gamma_{(Y)}$. The variance of this estimate consists of a

between-imputation component (B) and a within-imputation component (W):

$$\begin{aligned} \mathrm{B} =& \frac{1}{\mathrm{M}-1}\sum_{m=1}^{\mathrm{M}}\left(\gamma_{(Y)}^{m}-\overline{\gamma}_{(Y)}\right)^{2}, \\ \mathrm{W} =& \frac{1}{\mathrm{M}}\sum_{m=1}^{\mathrm{M}}\mathrm{V}(\gamma_{(Y)}^{m}), \end{aligned} \tag{4.9}$$

where $\mathrm{V}(\gamma_{(Y)}^{m})$ is the error variance connected with the fixed effect of interest based on the m*th* set of imputed values. The total error variance for the estimate $\overline{\gamma}_{(Y)}$ is

$$\mathrm{V}\left(\overline{\gamma}_{(Y)}\right) = \mathrm{W} + \frac{\mathrm{M}+1}{\mathrm{M}}\mathrm{B} \tag{4.10}$$

4.6 Results

Research Question 1: How similar are scores from the different scoring approaches? How much uncertainty do we see in the estimates of teacher practice?

There were two main classes of scoring approaches: (a) classroom aggregate means scores, and (b) scores based on the multilevel item bifactor measurement model (e.g., the EAP approach, the EAP with conditional means approach, and the plausible values approaches). In the first approach, scores were produced for each of the 7Cs separately, and therefore the scores were expected to be highly correlated. Table 4.5 presents the Pearson correlations among the 7Cs when classroom scores are produced using aggregated mean scores. Consistent with the findings of Raudenbush and Jean (2014), I found that the Tripod 7Cs dimensions are highly correlated when using this standard scoring approach. Therefore, I expect multi-collinearity issues when these scores are included simultaneously as classroom-level predictors in the hierarchical linear outcome model. By comparison, the scoring approaches that rely on the multilevel item bifactor measurement model are already accounting for these high inter-correlations by including a general factor that explains the shared variance among the dimensions. Table 4.6 presents the Pearson correlations among the 7Cs, where the scores are averaged plausible values drawn from the multilevel item bifactor analysis model with latent regression. It is clear that the use of the bifactor measurement model has produced

scores for the 7Cs that are not highly inter-correlated.

The scores under the two EAP approaches were found to be very highly correlated with each other (in the range of .85 to .99). Table 4.7 provides descriptive statistics for the scores from the plausible values and two EAP approaches. The descriptive statistics for the plausible values are based on averaging the sample descriptions (mean, SD, min, max) for each dimension across the 10 imputations. While the means for each dimension remain generally the same, the scores from the standard EAP approach (where all scores are shrunk to the same grand mean) have smaller standard deviations and smaller score ranges than the other two approaches. The difference between the two EAP scoring approaches pertained to the prior distribution. In the EAP with conditional means approach, the prior distribution $f(\boldsymbol{\vartheta}|\mathbf{X})$ that is used is conditional on observed covariates, whereas in the standard EAP approach, prior distribution $f(\boldsymbol{\vartheta})$ assumes the grand mean is identical for all of the classrooms in the sample. In the case of the Tripod survey, the likelihood distribution is based on all of the student item responses in a classroom, which is contributing more information to the posterior than the prior distribution. Therefore, it is not surprising that the factor scores from the two EAP approaches are so similar.

However, given the EAP scores are point estimates based on posterior means, the uncertainty (quantified by the standard errors) in the EAP scores is not carried through to the final outcomes model. This is problematic because in the multilevel outcomes model, the variance of the EAP scores will appear in the denominator of the estimate of the regression coefficient. This variance consists of both error and true variance, and inflates the denominator of the regression coefficients, causing the coefficients of interest (those pertaining to the teacher practice variables) to be attenuated.

Figure 4.2 presents the level of uncertainty in classroom-level Overall teacher effectiveness scores using the plausible values approach. Classroom scores on this dimension are sorted, and the plausible score ranges are plotted as a vertical line for each classroom. The red dashed horizontal lines represent the 25 and 75 percentile of scores. As seen in Figure 4.2, the level of uncertainty in the teacher practice scores differs by classroom, with some classrooms displaying a larger amount of variation within the

plausible values. The level of uncertainty is due to multiple factors, including classroom size and level of agreement among students regarding teacher practice. Figure 4.3 presents a similar set of plausible values intervals for the Challenge dimension. It is clear that the dispersion of plausible values for the classroom overall score is far less than that of the Challenge scores. This is due in part because the Challenge scale is based on eight items, whereas the Overall score is based on 36 items.

Figure 4.4 presents side-by-side boxplots of the 10 sets of imputed values for the Overall teacher effectiveness dimensions. These boxplots provide information about the distributions of plausible values and how similar the 10 sets are to each other. The distributions appear very similar across the 10 sets of plausible values, and large outliers are not observed in any of the sets.

Research Question 2: How do the resulting inferences about policy-relevant dimensions of teacher practices differ using standard class-aggregated mean scores from those based on a multilevel multidimensional plausible values approach?

Table 4.7 presents a comparison of the outcome model results across the four scoring approaches: (a) plausible values, (b) EAP scores with latent regression, (c) EAP scores without latent regression, and (d) standardized class mean scores. The first column of Table 4.7 displays the aggregated results from the outcome model that was estimated using 10 sets of plausible values. The reported estimates and standard errors are calculated following Equations 8 through 10. Using approach (a), the Overall teacher effectiveness dimension was found to be moderately related to end-of-year student achievement (B=.07, SE=.03, p<.05), controlling for the other predictors in the model. The Control (classroom management) was found to be a significant predictor of student achievement (B=.08, SE=.02, p<.0001), whereas Challenge was a modest predictor of achievement (B=.04, SE=.02, p<.05). The other 7Cs dimensions (Care, Clarify, Captivate, Confer, Consolidate) were not significantly related to student achievement.

The results are almost identical for the Plausible Values, EAP with conditional means, and EAP scoring approaches. The standard error for the Overall dimension in the plausible values approach is larger than the two EAP approaches, and the Challenge dimension for the first EAP approach is slightly larger (B=.05, SE=.02, p<.01). In general, the interpretations of the findings for these three approaches would be the

same.

The standardized class means score approach results in different regression coefficients for the 7Cs. Across all of the 7Cs predictors, the standard errors are larger than for the EAP and plausible values predictors, due to collinearity issues in the model. Control remains a significant predictor of student achievement (B=.08, SE=.03, p<.01), but now Challenge is also a significant predictor (B=.18, SE=.06, p<.01). The coefficients for Clarify and Confer are now negative, though they are not statistically significant.

4.7 Discussion

Many of the key policy questions in education involve relating unobserved (latent) characteristics of students, teachers, and schools (e.g., student engagement, teacher instructional practice, school climate) to student achievement. In these analyses, data are invariably multilevel, and the latent trait of interest is often measured by surveys with some degree of measurement error. The multilevel latent variable plausible values approach was proposed by Yang and Seltzer (2016), building off the multiple imputation methods used for large-scale assessments (see Mislevy et al., 1992), to appropriately handle measurement error in the predictors of a multilevel model. In this study, I expanded upon the multilevel latent variable plausible values approach to use a multilevel item bifactor analysis model as the measurement model for imputation. The choice to use a bifactor model is based on the fact that the Tripod survey conceptualizes teacher effectiveness as a general teacher property that is composed of seven interrelated domains (e.g., the 7Cs).

The multilevel multidimensional plausible values approach is outlined through an illustrative example of student surveys of teacher practice. Recent education policy has emphasized measures of teacher practice as an important lever for improving teacher quality and student achievement. Among these measures, student surveys of teacher practice have emerged as a popular and affordable choice for measuring what is occurring in the classroom. When I applied the multilevel multidimensional plausible

values approach to the Tripod data, I found that Overall teacher effectiveness was strongly related with student performance in English, controlling for key student and teacher background characteristics. Additionally, the Control (classroom management) domain provided unique prediction of student's academic achievement. That is to say, classrooms that were well behaved produced comparatively large learning gains. Other dimensions of practice that have been theorized to relate to student learning, such as whether a teacher is caring or able to captivate a student's interest, were not found to be related to student learning, holding constant teacher's overall level of effectiveness and other dimensions of practice.

There are multiple reasons why classroom management (Control) stands out as the only dimension of the 7Cs that is strongly predictive of student learning. There is a large body of literature that has identified the importance of classroom management skills as a precursor for learning (Brophy & Good, 1986; Emmer & Stough, 2001). Another reason specific to the Tripod survey is that the Control dimension has the highest intraclass correlation (percentage of variance between classrooms) of the dimensions and the largest distribution of classroom scores. That is to say, there is more information to distinguish among teachers' classroom management in the sample. Control items such as "Students in this class treat the teacher with respect" and "Student behavior in this class is a problem" have higher ICCs than items from any other domain, indicating that students are more able to agree on classroom behavior than other dimensions of the teacher's practice. Additionally, there is a much larger distribution of Control scores than the other 7Cs, indicating students are better able to differentiate between teachers with high and low levels of classroom management skills.

The other central finding from this study is that in the context of relating Tripod scores and student learning, I see very little gain from using the plausible values approach over more widely-used EAP scores. This finding differs from Yang and Seltzer (2015), who found large gains in terms of the magnitude of regression coefficients in the outcomes model from the plausible values method compared with EAP or summed scores. In the Yang and Seltzer study, scores were imputed for teachers (level-1) and

schools (level-2) based on a three item measure of teacher practice. In my study, scores are estimated for classrooms (level-2) based on student (level-1) responses on a 36 item scale, where there are a minimum of three items per dimension. In my sample, there are on average 16 students per class with a total of 18,000 students, and so I am able to get both (a) precise estimates of the item and structural parameters in the measurement model, and (b) precise EAP scores for classrooms where the amount of shrinkage to conditional means or the grand mean is minimal.

Conclusions regarding the predictive validity of the Tripod measure are limited by the fact that the MET sample is a non-representative sample of students in districts, as schools and teachers volunteered to participate in this research study. There is also a fair amount of missing data due to inconsistencies in administrative records, which further limits the representativeness of the sample. Additionally, in the imputation scoring model, the parameters are fixed to MLE estimates, so the imputations do not reflect all of the uncertainty from the model. A fully Bayesian approach could be estimated to account for uncertainty in model parameters in the distribution posterior, which might lead to a larger differentiation between the EAP and plausible values scoring approaches.

Further work will be conducted to provide better insight into the conditions under which the multilevel multidimensional plausible values approach improves upon the inferences that can be drawn using simpler measurement approaches. It is likely that the usefulness of this approach will depend on the interaction between the number of items in the measurement model, the number of level-1 and level-2 sample sizes, whether the latent variables of interest are at the individual or group level, and the degree of collinearity among predictors in the imputation model.

Table 4.1: Student, class, and teacher characteristics

Category Name	M /%	SD	ICC
Outcome			
End of year ELA score (spring 2010)	0.17	0.94	0.26
Student (N=13,989)			
Male	0.49	0.50	0.01
Gifted	0.13	0.34	0.24
English learner (EL)	0.13	0.34	0.14
White	0.23	0.44	0.35
Black	0.26	0.45	0.49
Hispanic	0.41	0.48	0.36
Asian	0.08	0.27	0.34
Special education	0.05	0.22	0.24
Free or reduced price lunch (FRPL)	0.62	0.49	0.34
Prior year ELA score (spring 2009)	0.16	0.94	0.26
Classroom (N=884)			
Male	0.50	0.14	—
Gifted	0.12	0.21	—
English Learner (EL)	0.14	0.18	—
White	0.23	0.30	—
Black	0.28	0.28	—
Hispanic	0.41	0.29	—
Asian	0.06	0.11	—
Free or reduced price lunch (FRPL)	0.64	0.31	—
Special education	0.07	0.12	—
Prior year ELA test score	0.09	0.66	—
Number of students per class	16.57	5.69	—
Teacher (N= 463)			
Years of experience			
0 to 3 years	0.27	0.44	—
4 to 6 years	0.19	0.39	—
7 or more years	0.55	0.50	—
Male	0.15	0.36	—
Masters degree or higher	0.29	0.45	—

Note. ICC = intraclass correlation, and refers here to the percentage of variance between schools.

Table 4.2: Descriptive statistics for the Tripod student perceptions survey

Domain	Item	Item Wording	Mean	SD	ICC
Care	A10	My teacher in this class makes me feel that s/he really cares about me.	3.71	1.23	0.26
Care	B146	My teacher seems to know if something is bothering me.	3.12	1.31	0.25
Care	B34	My teacher really tries to understand how students feel about things.	3.57	1.19	0.25
Control	B112	Student behavior in this class is under control.	3.36	1.24	0.25
Control	B113*	I hate the way that students behave in this class.	3.50	1.32	0.23
Control	B114*	Student behavior in this class makes the teacher angry.	2.98	1.30	0.27
Control	B138*	Student behavior in this class is a problem.	3.31	1.27	0.28
Control	B46	My classmates behave the way my teacher wants them to.	3.13	1.20	0.28
Control	B49	Students in this class treat the teacher with respect.	3.59	1.13	0.32
Control	B6	Our class stays busy and does not waste time.	3.46	1.15	0.23
Clarify	B1	If you don't understand something, my teacher explains it another way.	4.04	1.04	0.20
Clarify	B130	My teacher knows when the class understands, and when we do not.	3.81	1.08	0.18
Clarify	B136	When s/he is teaching us, my teacher thinks we understand even when we don't.	3.60	1.20	0.18
Clarify	B17	My teacher has several good ways to explain each topic that we cover in this class.	3.92	1.04	0.24
Clarify	B80	My teacher explains difficult things clearly.	3.86	1.07	0.21
Challenge	B128	My teacher asks questions to be sure we are following along when s/he is teaching.	4.29	0.96	0.19
Challenge	B133	My teacher asks students to explain more about answers they give.	4.06	0.98	0.20
Challenge	B21	In this class, my teacher accepts nothing less than our full effort.	3.97	1.06	0.20
Challenge	B36	My teacher doesn't let people give up when the work gets hard.	4.00	1.09	0.20
Challenge	B45	My teacher wants us to use our thinking skills, not just memorize things.	4.10	1.01	0.17

Challenge	B59	My teacher wants me to explain my answers–why I think what I think.	4.07	1.01	0.21
Challenge	B70	In this class, we learn a lot almost every day.	3.79	1.09	0.22
Challenge	B90	In this class, we learn to correct our mistakes.	4.00	1.03	0.22
Captivate	B141*	This class does not keep my attention–I get bored.	3.41	1.33	0.19
Captivate	B29	My teacher makes learning enjoyable.	3.62	1.24	0.30
Captivate	B44	My teacher makes lessons interesting.	3.59	1.21	0.29
Captivate	B89	I like the ways we learn in this class.	3.83	1.00	0.26
Confer	B129	My teacher wants us to share our thoughts.	3.89	1.11	0.24
Confer	B135	Students get to decide how activities are done in this class.	2.45	1.09	0.23
Confer	B154	My teacher gives us time to explain our ideas.	3.73	1.10	0.26
Confer	B155	Students speak up and share their ideas about class work.	3.60	1.16	0.20
Confer	A54	My teacher respects my ideas and suggestions.	3.79	1.12	0.23
Consolidate	B145	My teacher takes the time to summarize what we learn each day.	3.50	1.20	0.22
Consolidate	B147	My teacher checks to make sure we understand what s/he is teaching us.	4.01	1.06	0.24
Consolidate	B58	We get helpful comments to let us know what we did wrong on assignments.	3.77	1.15	0.21
Consolidate	B83	The comments that I get on my work in this class help me understand how to improve.	3.81	1.13	0.21

Note. ICC = intraclass correlation, and refers here to the percentage of variance between classrooms.

*Item is reverse-coded.

Table 4.3: Item parameter estimates for the multilevel measurement model

Item	a1	a2	a3	a4	a5	a6	a7	a8	a9
A10	1.90(.02)	.43(.01)	0(—)	0(—)	0(—)	0(—)	0(—)	0(—)	1.90(.02)
B146	1.36(.02)	.43(.01)	0(—)	0(—)	0(—)	0(—)	0(—)	0(—)	1.36(.02)
B34	2.03(.02)	.43(.01)	0(—)	0(—)	0(—)	0(—)	0(—)	0(—)	2.03(.02)
B112	0.86(.02)	0(—)	.81(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	0.86(.02)
B113	0.21(.02)	0(—)	.81(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	0.21(.02)
B114	0.53(.02)	0(—)	.81(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	0.53(.02)
B138	0.50(.02)	0(—)	.81(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	0.50(.02)
B46	1.03(.02)	0(—)	.81(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	1.03(.02)
B49	1.09(.02)	0(—)	.81(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	1.09(.02)
B6	0.99(.02)	0(—)	.81(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	0.99(.02)
B1	1.88(.02)	0(—)	0(—)	.23(.01)	0(—)	0(—)	0(—)	0(—)	1.88(.02)
B130	1.46(.02)	0(—)	0(—)	.23(.01)	0(—)	0(—)	0(—)	0(—)	1.46(.02)
B136	0.76(.02)	0(—)	0(—)	.23(.01)	0(—)	0(—)	0(—)	0(—)	0.76(.02)
B17	2.13(.03)	0(—)	0(—)	.23(.01)	0(—)	0(—)	0(—)	0(—)	2.13(.03)
B80	2.06(.02)	0(—)	0(—)	.23(.01)	0(—)	0(—)	0(—)	0(—)	2.06(.02)
B128	1.28(.02)	0(—)	0(—)	0(—)	.32(.01)	0(—)	0(—)	0(—)	1.28(.02)
B133	1.04(.02)	0(—)	0(—)	0(—)	.32(.01)	0(—)	0(—)	0(—)	1.04(.02)
B21	1.27(.02)	0(—)	0(—)	0(—)	.32(.01)	0(—)	0(—)	0(—)	1.27(.02)
B36	1.58(.02)	0(—)	0(—)	0(—)	.32(.01)	0(—)	0(—)	0(—)	1.58(.02)
B45	1.45(.02)	0(—)	0(—)	0(—)	.32(.01)	0(—)	0(—)	0(—)	1.45(.02)
B59	1.35(.02)	0(—)	0(—)	0(—)	.32(.01)	0(—)	0(—)	0(—)	1.35(.02)
B70	1.73(.02)	0(—)	0(—)	0(—)	.32(.01)	0(—)	0(—)	0(—)	1.73(.02)
B90	1.91(.02)	0(—)	0(—)	0(—)	.32(.01)	0(—)	0(—)	0(—)	1.91(.02)
B141	1.11(.02)	0(—)	0(—)	0(—)	0(—)	.55(.01)	0(—)	0(—)	1.11(.02)
B29	2.09(.02)	0(—)	0(—)	0(—)	0(—)	.55(.01)	0(—)	0(—)	2.09(.02)
B44	2.19(.03)	0(—)	0(—)	0(—)	0(—)	.55(.01)	0(—)	0(—)	2.19(.03)
B89	1.62(.02)	0(—)	0(—)	0(—)	0(—)	.55(.01)	0(—)	0(—)	1.62(.02)
B129	1.09(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	.39(.01)	0(—)	1.09(.02)
B135	0.71(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	.39(.01)	0(—)	0.71(.02)
B154	1.98(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	.39(.01)	0(—)	1.98(.02)
B155	1.36(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	.39(.01)	0(—)	1.36(.02)
A54	2.06(.03)	0(—)	0(—)	0(—)	0(—)	0(—)	.39(.01)	0(—)	2.06(.03)
B145	1.54(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	0(—)	.25(.01)	1.54(.02)
B147	2.37(.03)	0(—)	0(—)	0(—)	0(—)	0(—)	0(—)	.25(.01)	2.37(.03)
B58	1.72(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	0(—)	.25(.01)	1.72(.02)
B83	1.76(.02)	0(—)	0(—)	0(—)	0(—)	0(—)	0(—)	.25(.01)	1.76(.02)
Mean	0(—)	0(—)	0(—)	0(—)	0(—)	0(—)	0(—)	0(—)	0(—)
Var.	.30(.02)	1(—)	1(—)	1(—)	1(—)	1(—)	1(—)	1(—)	1(—)

Note. Estimated Standard errors are in parentheses. Fixed parameters do not have estimated standard errors.

Table 4.4: Parameter estimates for the multilevel measurement model latent regression

	Overall (class)		Overall (student)	
Covariate	Mean	SD	Mean	SD
Student level fixed effects (group mean centered)				
Current year test score (spring 2010)	—	—	0.00	0.01
Prior year test score (spring 2009)	—	—	0.01	0.00
Gifted	—	—	0.02	0.02
White	—	—	-0.04	0.02
Free Reduced Price Lunch (FRPL)	—	—	0.01	0.00
Classroom level fixed effects (grand mean centered)				
Class average current year test score (spring 2010)	0.25	0.07	—	—
Class average prior year test score (spring 2009)	-0.09	0.05	—	—
% Gifted	-0.05	0.11	—	—
% White	-0.09	0.08	—	—
% FRPL	0.00	0.00	—	—
% Special Education	0.81	0.17	—	—
% English learner	0.34	0.11	—	—

Table 4.5: Pearson correlations among the class-aggregated scores for the 7Cs

	Care	Control	Clarify	Challenge	Captivate	Confer	Consol.
Care	1.00						
Control	0.54	1.00					
Clarify	0.83	0.58	1.00				
Challenge	0.80	0.65	0.88	1.00			
Captivate	0.85	0.64	0.83	0.79	1.00		
Confer	0.87	0.60	0.84	0.82	0.85	1.00	
Consolidate	0.85	0.62	0.88	0.89	0.84	0.86	1.00

Table 4.6: Pearson correlations among averaged plausible values from the multilevel measurement model

	Overall	Care	Control	Clarify	Chall.	Capt.	Confer	Consol.
Overall	1.00							
Care	-0.04	1.00						
Control	0.20	-0.16	1.00					
Clarify	0.00	-0.02	0.03	1.00				
Chall.	-0.11	-0.08	0.10	0.08	1.00			
Capt.	0.10	0.07	-0.06	-0.01	-0.18	1.00		
Confer	0.00	0.08	0.00	-0.04	0.00	-0.08	1.00	
Consol.	-0.10	0.03	-0.08	0.04	0.06	0.01	0.09	1.00

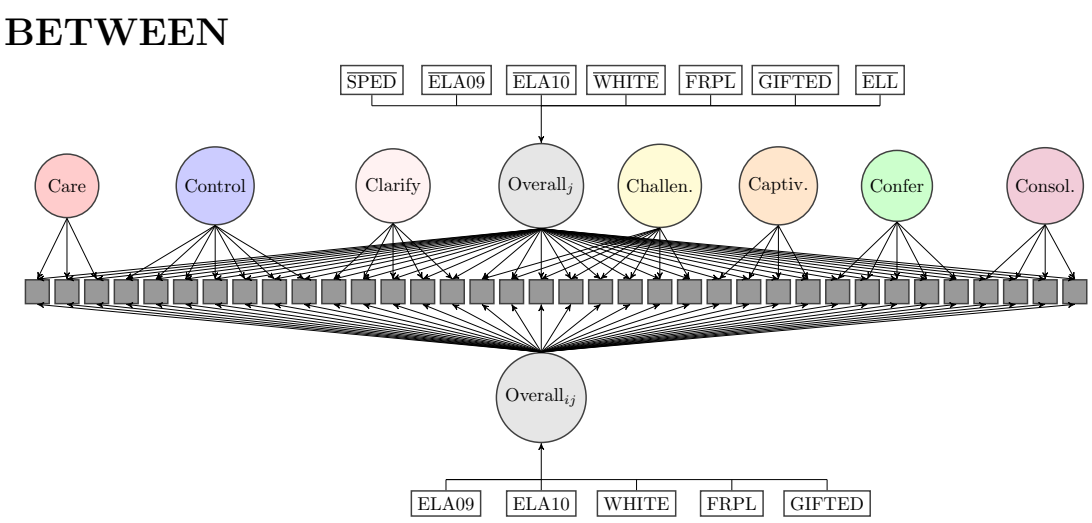
Note. These correlation from the bifactor measurement model where the specific factors are orthogonal to the Overall dimension.

Table 4.7: Results from the outcome model using the four scoring approaches

	Plausible Values	EAP (with latent regression)	EAP (w/o latent regression)	Mean Scores
(Intercept)	0.03(0.02)*	0.03(0.02)*	0.03(0.02)*	0.04(0.02)**
Prior year test score	0.64(0.01)***	0.64(0.00)***	0.64(0.01)***	0.65(0.01)***
Gifted	0.21(0.02)***	0.21(0.00)***	0.21(0.02)***	0.21(0.02)***
White	0.05(0.02)**	0.05(0.00)**	0.05(0.02)**	0.06(0.02)***
FRPL	-0.06(0.01)***	-0.06(0.00)***	-0.06(0.01)***	-0.06(0.01)***
Class avg. prior year test score	0.61(0.02)***	0.60(0.00)***	0.60(0.02)***	0.61(0.02)***
% Gifted	0.24(0.04)***	0.24(0.00)***	0.24(0.04)*	0.24(0.04)***
% White	0.07(0.04)	0.07(0.00)	0.07(0.04)	0.07(0.04)*
% FRPL	-0.03(0.04)	-0.03(0.00)	-0.03(0.04)	-0.03(0.04)
% Special Education	-0.10(0.07)	-0.10(0.00)	-0.10(0.07)	-0.10(0.07)
% ELL	0.03(0.05)	0.03(0.00)	0.03(0.05)	0.03(0.05)
Overall	0.07(0.03)*	0.07(0.00)**	0.07(0.02)**	—(—)
Care	-0.01(0.02)	-0.02(0.00)	-0.02(0.02)	0.00(0.05)
Control	0.08(0.02)***	0.08(0.00)***	0.08(0.01)***	0.08(0.03)**
Clarify	0.00(0.02)	0.00(0.00)	0.00(0.02)	-0.14(0.08)
Challenge	0.04(0.02)	0.05(0.00)**	0.03(0.02)*	0.18(0.06)**
Captivate	-0.01(0.02)	-0.01(0.00)	-0.02(0.02)	0.02(0.05)
Confer	0.00(0.02)	0.01(0.00)	0.01(0.02)	-0.07(0.06)
Consolidate	-0.01(0.02)	-0.02(0.00)	-0.01(0.02)	0.00(0.06)

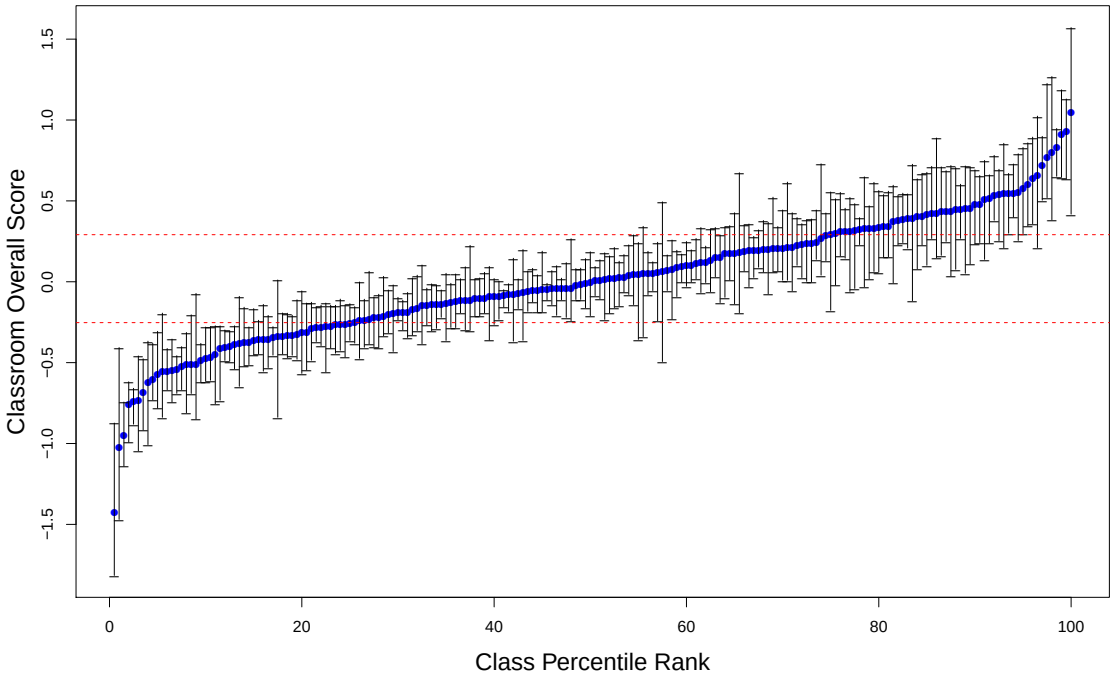
Note. All models included district fixed effects. Estimates are suppressed due to MET data requirements.
*p<.05. **p<.01. ***p<.001

Figure 4.1: Path diagram of the stage 1 multilevel bifactor measurement model



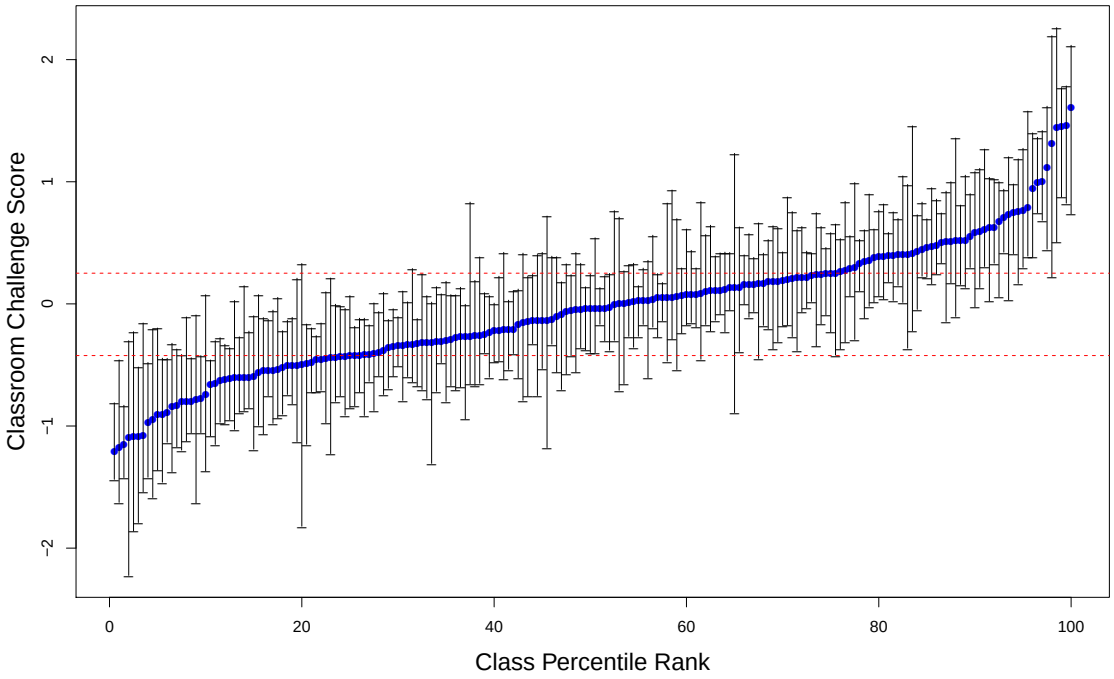
This figure shows the measurement model with latent regression. The variable names are defined in Section 4.5.3.

Figure 4.2: Caterpillar plot of classroom Overall teacher effectiveness plausible value scores, sorted by class ranking



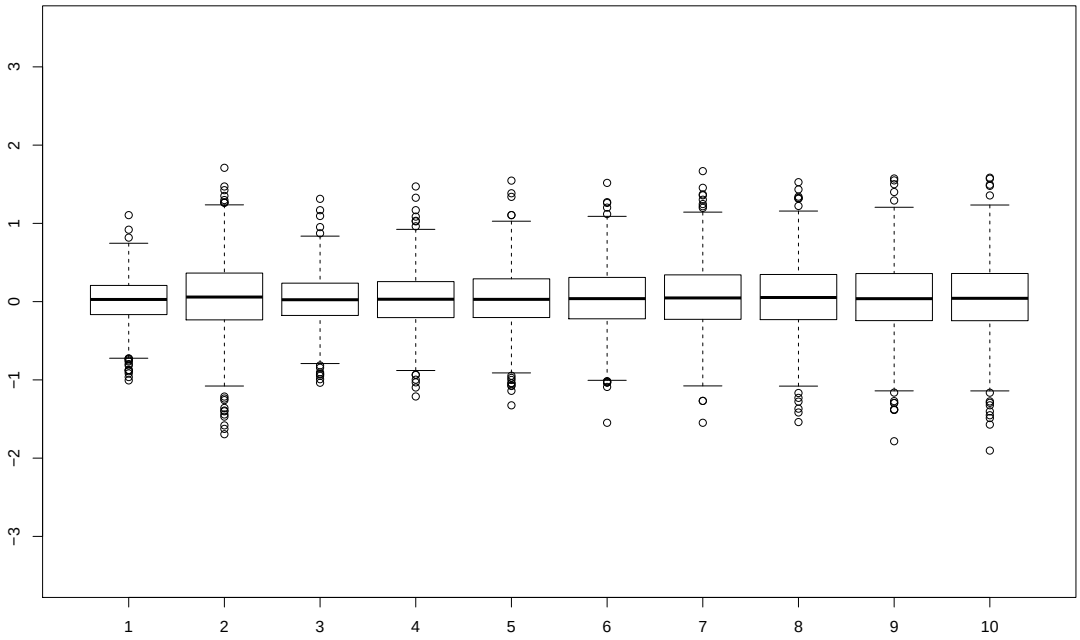
NOTE: This plot displays the classroom-level plausible values for the overall teacher effectiveness for a random sub-sample of 200 classrooms. This sub-sample, rather than the full sample, is displayed in the figure because the full sample could not be neatly fit into a single plot. The average plausible value score for each classroom is plotted as a blue circle. The black horizontal lines represent the minimum and maximum plausible value draw for each classroom, and vertical line represents the range of the interval. The red dashed horizontal lines represent the 25 and 75 percentile of scores.

Figure 4.3: Caterpillar plot of classroom Challenge plausible value scores, sorted by class ranking



NOTE: This plot displays the classroom-level plausible values for the overall teacher effectiveness for a random sub-sample of 200 classrooms. This sub-sample, rather than the full sample, is displayed in the figure because the full sample could not be neatly fit into a single plot. The average plausible value score for each classroom is plotted as a blue circle. The black horizontal lines represent the minimum and maximum plausible value draw for each classroom, and vertical line represents the range of the interval. The red dashed horizontal lines represent the 25 and 75 percentile of scores.

Figure 4.4: Boxplot of the 10 sets of imputed values for the Overall teacher effectiveness dimension



CHAPTER 5

A validity analysis of the Tripod survey

5.1 Abstract

This article develops a validity argument for the use of student surveys of instructional practice to assess teacher effectiveness in summative teacher evaluations and professional development decisions. The Tripod Survey, a student perceptions survey that has been administered to over 100,000 classrooms in the United States, is reported to measure student ratings of seven theoretical dimensions (e.g., the 7Cs^{TM}) of teacher effectiveness. Despite its use in teacher evaluation systems across the country, very little in-depth psychometric analysis has been published on the Tripod survey. Using data collected by the Measures of Effective Teaching (MET) Project in six large U.S. school districts, I build a validity argument for the use of the Tripod Survey to measure teacher practice in middle school English and math classrooms. These analyses found that Tripod scores are fairly reliable and correlated with classroom observation ratings and teacher value-added scores. However, caution is suggested in interpreting these results as an endorsement for use in high-stakes teacher evaluations.

5.2 Introduction

Recent federal policies have catalyzed changes in teacher evaluation systems across the country. Race to the Top (RTTT), a competitive grant program that begun in 2009, called for states who were competing for millions of dollars' worth of funding to establish more rigorous teacher evaluation systems that rely on multiple measures (U.S. Department of Education, 2009). Additionally, in 2011, a program was started to provide waivers to the Elementary and Secondary Education Act (ESEA). In order to

get a waiver, state education agencies were encouraged to develop teacher evaluation systems that emphasized the use of multiple measures, with student growth used as a significant factor in the evaluation system (Popham, 2013). According to the National Council of Teacher Quality, between 2009 and 2012, 36 states and the District of Columbia introduced new teacher evaluation policies (Doherty & Jacobs, 2013).

One of the measures that is being increasingly introduced in new teacher evaluation systems is student perception surveys of instructional practice. Seven states now mandate that student surveys are required as a component of teacher evaluations, while 26 other states allow for the use of students surveys in teacher evaluations (Doherty & Jacobs, 2015). These surveys ask students about their opinions about specific teachers and specific classrooms. The goal of these surveys is to provide fair and reliable feedback to teachers regarding their students' perceptions of the strengths and weaknesses of a range of teacher practices, including academic rigor, classroom management skills, and academic support. Additionally, surveys are being used to guide professional development programs (Bill & Melinda Gates Foundation, 2012).

Proponents of student surveys note that students are natural observers of the classroom in which they spend their days, and provide feedback that adult observers cannot get a single classroom observation (Ferguson, 2012). The main advantages cited by survey proponents are that survey results point to strengths and areas for improvement, the items have face validity and reflect what teachers value, and survey results demonstrate relatively high consistency (Bill & Melinda Gates Foundation, 2012).

The central concern raised about student surveys is that students may not be objective raters of their teachers (Marsh, 1987; Theall & Franklin, 2001; Liaw & Goh, 2003). As researchers from the Measures of Effective Teaching Project have acknowledged, "although most of the concern regarding bias has focused on the achievement gain measures, the non-test-based components could also be biased by the same unmeasured student traits that would cause bias with the test-based measures" (Kane et al., 2013, pg. 6). Following the definition used in Centra (2003, p. 498), bias exists when a "student, teacher, or course characteristic affects the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching, such as in-

creased student learning." As acknowledged by Bell et al. (2012), teacher effectiveness and contextual features of the classroom are intertwined, and therefore instruments that measure teacher effectiveness should investigate the instrument's sensitivity to contextual features.

This study focuses on the Tripod Survey (Ferguson, 2010), which is the most widely-used off-the-shelf student survey instrument (Bill & Melinda Gates Foundation, 2012). The Tripod student perceptions assessment was developed by Ron Ferguson at Harvard University, and is based upon classroom-level surveys developed by the Tripod Project for School Improvement (Ferguson, 2010). The "tripod" describes the knowledge and skills that are needed to deliver instruction effectively: (a) content knowledge, (b) pedagogic knowledge and skills, and (c) the ability to connect with students on a personal level. The Tripod survey focuses primarily on what teachers do and how the classroom operates, which is operationalized as the Tripod 7Cs framework of teacher effectiveness. The seven scales are: Care, Control, Clarify, Challenge, Captivate, Confer and Consolidate, which are described in more detail later in the paper. The developers state that the "Tripod survey results provide information that teachers can use to set specific priorities for differentiated professional development and coaching support" (Tripod Project, 2016). Additionally, some districts and states have decided to include scores from the Tripod survey as a weighted component of an overall teacher evaluation plan (see Table 2.1 Schweig, 2014)

The purpose of this paper is to explore the validity of the claims about teacher effectiveness that can be made using the Tripod secondary survey. Kane (2006) describes the purpose of validity research as articulating an integrated argument to describe the degree to which an instrument has been validated for a particular purpose. Validity arguments have been previously developed for the use of value-added models (Haertel, 2013) and classroom observation protocols (Bell et al., 2012) in high-stakes personnel decisions. Despite the recent wide-spread use of student surveys in teacher evaluation systems across the country, very little validation work has been published on the Tripod survey. This paper represents an advancement from current research for three reasons: (a) it draws from current validity thinking to examine the evidence for sepa-

rate purposes (in summative and formative teacher evaluations) of the Tripod survey, (b) it takes advantage of a large, extant dataset that includes multiple measures of teacher effectiveness, and (c) it relies on methodological advances to make inferences regarding student surveys accounting for both the multilevel nature of school data and contextual classroom features.

In the following sections, I describe an interpretive argument approach using Michael Kane's (2006) validity argument approach. I focus on two purposes of the Tripod survey—use in a summative high-stakes teacher evaluation system and use for professional development decisions. I begin by describing the Kane framework for validity. Then, I apply this validity framework to student survey data collected by the Measures of Effective Teaching (MET) Project in English and math middle school classrooms during the 2009–2011 school years in six U.S. school districts. Lastly, I reflect on the use of this validity approach with student survey data and the remaining unresolved validity issues regarding the high-stakes use of surveys.

5.3 The validity argument approach

Validity research is concerned with the degree to which an instrument has been validated for a particular purpose. In the analysis of the student ratings of teacher practice, I rely on a validity argument framework to investigate the case for using the Tripod student perceptions survey to measure the effectiveness of teachers. Michael Kane's (2006) "Validation" chapter serves as the framework for making a validity argument. In this framework, an *interpretive argument* is first made, which discusses "the inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances" (Kane, 2006, pg. 23). Secondly, the *validity argument* evaluates the level of empirical data to support each of these inferences and assumptions, focusing on the proposed uses of the scores. The four central steps in the interpretive argument with regards to the Tripod survey are outlined in Table 5.1 and are described in more detail below.

5.4 The interpretive argument for the Tripod survey

5.4.1 Scoring

The first step is broadly called scoring, but actually contains multiple important components for making claims based on the Tripod Survey. The scoring argument holds that teacher scores produced from the Tripod survey accurately capture dimensions of instructional quality undistorted by other factors. The principal concerns in this step are the degree to which scores are reflective of the dimensions of the survey, the degree of bias in scores, and the appropriateness of the scoring model.

Dimensionality. The first stage of the scoring inference is to test the theoretical dimensionality of the Tripod survey. The survey developer states that items map onto seven dimensions of teaching practice, which are referred to as the 7Cs: Care, Control, Clarify, Challenge, Captivate, Confer and Consolidate (Ferguson, 2012). These domains are defined in Table 5.2. Additionally, these dimensions have been further grouped into two categories of practice: Press and Academic Support. Press is defined as "keeping students busy and on task and pressing them to think rigorously and persist in the face of difficulty" (Ferguson & Danielson, 2014, pg. 100), and contains the Challenge and Control domains. Academic support is defined by "caring teacher-student relationships, captivating lessons, and other practices that students experience as supportive", and contains the other five domains (Ferguson & Danielson, 2014, pg. 100). However, little psychometric work has been conducted to verify either the 7Cs or the press and support dimensionality structure. The one reported analyses that I am aware of was conducted by Ferguson (2010), who performed a factor analysis with class averages of student survey responses and found the Control index was distinguished from the other six Cs.

Bias. The second component of the scoring inference is to check for score bias. Popham (2013) outlines several sources of rater bias that may be a concern, including: (a) *severity error* – a rater's predisposition to supply lower ratings independent of the content being rated, (b) *generosity error* – the opposite of severity error, where the rater will be more likely to use higher response categories, and (c) *central-tendency*

error – the predisposition to use the middle rating category. A whole line of research regarding response styles has been developed to examine the impact of these types of rater bias on scores (e.g., De Jong, Steenkamp, Fox, & Baumgartner, 2008; Falk & Cai, 2015), but this work has not so far been applied to student ratings of instructional practice.

Scoring Model. The last component of the scoring inference is to examine how scores are produced. Every method of score aggregation implies a scoring model, which can be examined in terms of data fit. Tripod scores are typically created by averaging student responses to items to the classroom level. Scores have also been reported in terms of degree of agreement (percentage of student responding Mostly True or Totally True) at the item and domain level. Furthermore, scores have been produced for the MET study using multiple regression to adjust scores to account for student characteristics and student baseline test scores (Kane et al., 2013). The residuals from this regression form the adjusted classroom level student perception survey scores that are reported in the MET study datasets. The fit of the scoring models can be examined by looking at the similarity of scores across scoring models, as well as the how well various scores predict key outcomes of interest (e.g., student test scores).

5.4.2 Generalization

The second step is generalization, which is concerned with the reliability and stability of scores. If the intention is to use the Tripod scores for personnel decisions, the generalization inference might be to all the courses the teacher taught that year. If evaluations happen on a biennial basis, it is necessary to establish the consistency of the Tripod scores for a given teacher across multiple school years. Therefore, the principal focus of the generalization inference is the consistency of Tripod scores across different class sections of as well as over time for a given teacher.

Reliability. The first component of this step is to examine the marginal reliability of the Tripod item sets, as well as to examine within-group agreement in the ratings of teachers within a classroom.

Stability. The second component of generalization is to examine the stability of

scores across sections within a year and across years. The across-section and across-year correlation of Tripod scores within a teacher is examined in this paper to estimate the proportion of observed variation that is due to stable differences between teachers.

5.4.3 Extrapolation

The goal of the extrapolation argument is to measure how strongly student-reported measures of teacher effectiveness are related to a broader understanding of teacher effectiveness. The first part of the extrapolation analyses in this study is to replicate the correlational analysis in Kane and Cantrell (2010) and Kane and Staiger (2012) examining the relationship between classroom Tripod scores, value-added model (VAM) scores, and classroom observation scores. However, in the correlational analyses presented here, I am using latent teacher practice scores from a multilevel item factor analysis (IFA) measurement model (Kuhfeld et al., in preparation). I also expand upon the previous analyses by looking at other measures of teacher practice and classroom environment than those previously examined by Kane and Cantrell (2010), including student happiness in the classroom and teacher content knowledge. The second part of the extrapolation argument is to relate end-of-year student achievement and the Tripod dimensions. The original validity question proposed by MET was whether "any additional components of the evaluation (e.g., classroom observations, student feedback) should be demonstrably related to student achievement gains" (Kane & Cantrell, 2010, p. 5). While others have predicted teacher value-added with classroom-level Tripod scores (see Raudenbush & Jean, 2014; Ferguson & Danielson, 2014), little work has been done to directly relate student-level achievement outcomes and the Tripod domains. This study uses the Tripod dimensions as predictors of student academic achievement in a hierarchical linear model, controlling for student and classroom characteristics.

5.4.4 Implication

Implication is the final step in the interpretive argument, and this step is centrally concerned with how scores will be used by various stakeholders. The same measure can be valid for one purpose and invalid for another. In this step, I outline the specific

proposed uses of the Tripod survey that have been discussed in the literature or proposed by districts and states, and then review the evidence that supports or negates each use. Specifically, the Tripod survey has been used previously as a weighted component of a summative teacher evaluation, as well as to provide formative feedback that can guide professional development and coaching efforts. Furthermore, potential unintended consequences of using these scores to measure teacher effectiveness are considered.

5.5 The validity argument for the Tripod survey

For the validity argument, I use data from the Measures of Effective Teaching (MET) project, which is the largest study of classroom teaching ever conducted in the United States. Data were collected on a variety of measures of teacher quality over two academic school years, 2009–2010 and 2010–2011, within six large school districts in the United States. More than 2,500 fourth- through ninth-grade teachers working in 317 schools participated in the study (White & Rowan, 2013).

5.5.1 Sample

The purpose of Measures of Effective Teaching (MET) Project was to examine the reliability and validity of the various measures of teacher effectiveness. Teacher volunteers were selected from six participating districts: Charlotte-Mecklenburg (North Carolina) Schools, Dallas (Texas) Independent School District, Denver (Colorado) Public Schools, Hillsborough County (Florida) Public Schools, Memphis (Tennessee) City Schools, and the New York City (New York) Department of Education (White & Rowan, 2013).

I focus on students and teachers in Grades 6 to 8 in English and math classrooms. The MET data also contains elementary students, who filled out an elementary version of the Tripod survey. The properties of the elementary survey will be examined in future work.

Table 5.3 summarizes the student, classroom, and teacher characteristics across both years of the MET study in math and English middle school classrooms. The

student Tripod responses in the 2009–10 school year were used as the calibration sample for the Scoring and Extrapolation analyses. The total English sample in 2009–10 included 19,245 students, who were nested within 1,071 class sections taught by 572 teachers. The math sample in the first year of the study included 16,716 students, who were nested within 907 class sections taught by 494 teachers. The student Tripod responses collected in 2010–11 were used in the Generalization step to measure stability of scores across years. The student sample in 2010–11 is smaller than 2009–10, and consists of the students of the sub-sample of teachers that were in both years of the study.

5.5.2 Measures

This study focuses on the student responses to the secondary Tripod survey, which was administered in both years of the study. The Tripod survey asked students to rate many aspects of teacher's practices and behavior towards students. There are 36 Tripod items measuring teacher practice in the secondary survey. A Likert-type response scale with a 5 response options (Totally Untrue; Mostly Untrue; Somewhat; Mostly True; Totally True) was used. Table 5.4 provides the wording of the items and the 7Cs dimension that each item is hypothesized to measure, along with the mean and standard deviation of item responses for the students in middle school English and math classrooms in 2009–10.

Tripod surveys were administered both on paper and online, with the participating schools choosing the mode of administration. Paper surveys were distributed to students with their names on peel-off labels that they removed before completing them, so that no school personnel could use to identify respondents when surveys were completed. Study-specific barcodes remained on the survey to allow for the linking of student responses to child characteristics. Similar verification procedures were used for online administration, but with teachers distributing login codes with unique identifiers (Bill & Melinda Gates Foundation, 2012).

As a part of the validity analysis, the student achievement measures, value-added scores, teacher content knowledge, and classroom observation ratings will also be used.

These measures are described below.

Student achievement. At grades 6-8, state end-of-year standardized assessments administered in each district (typically in reading and mathematics) were used to measure student achievement. All student achievement scores were first converted to rank-based z-scores within district, subject, and grade. Both sets of student test scores were used to create classroom section value-added model (VAM) scores (see White and Rowan (2013) for a description of the VAM model).

Student-reported measures. In addition to the Tripod student perception items, the student survey contained scales measuring effort exerted in class, happiness in class, and the amount of test prep activities in the class. During the 2009-10 school year, students also reported on their college aspirations and how often they read at home. Student scale scores were created by taking the simple mean of the scale items. Class-level scores were estimated by aggregating student scale scores using simple means.

Teacher Pedagogical and Content Knowledge. The Content Knowledge for Teaching Assessment (CKT) was administered in the second year of the MET study. Separate assessments were created and administered for grades 4-6 ELA, grades 7-9 ELA, and grades 6-8 Mathematics.

Observation scores. A variety of observational protocols were used to assess classroom quality based on a set of video recordings of classroom lessons. Each teacher filmed multiple lessons teaching different topics, with one camera focused on the board and other providing a 360 degree classroom view. The Classroom Assessment Scoring System (CLASS; Pianta et al., 2008) and the Framework for Teaching (Danielson, 2007) are two of the more commonly-used observational protocols that were included in the MET study. The CLASS measure assumes 10 dimensions of teaching practice (grouped under four categories: Emotional Support, Instructional Support, Classroom Organization, and Student Engagement), while the Framework for Teaching assumes that there are eight dimensions of teacher practice (grouped under two categories: Classroom Environment and Instruction). White and Rowan (2013) describe the rater training and scoring procedures in detail. In general, raters scored classrooms using each observation protocol focusing on 15-30 minute video segments, and class-level

scores are produced by averaging raters' scores to get a single segment score and then calculating the harmonic mean across segments for a particular target of measurement.

5.6 Results

5.6.1 Scoring

As an exploratory step, I first fit a series of exploratory item factor analysis (EFA) models (Cai, 2010a) with the secondary Tripod survey data, extracting up to four latent dimensions. These models were fit to the student item responses, ignoring the nesting of students of in classrooms. An oblique rotation method (oblique CF-Quartimax rotation, described by Browne (2001)) was used to allow for the correlation of factors. Item factor analysis models were fit independently to the English and math datasets. The exploratory item factor analysis models were calibrated using full-information maximum likelihood estimation in flexMIRT® (Cai, 2015). As an additional source of information, I also fit a unidimensional IRT model and examined the standardized Chen and Thissen (1997) local dependence χ^2 indices. These local dependence indices can be used to detect residual associations between items, which may imply the presence of unmodeled dimensions. These first exploratory steps ignore the multilevel nature of the data, but can still provide important information regarding the dimensionality of the data.

Results from the single-level EFA are available in the supplemental materials. A unidimensional model appears to fit well, with all but three items (B113, B114, B138) displaying a standardized factor loading above 0.40. For both subjects, the set of items that Ferguson (2010) defines as Control (B112, B113, B114, B138, B46, B49, and B6) formed a second factor in the two-dimensional and three-dimensional EFA models. These items are primarily focused on student behavior in the classroom, whereas the focus of the majority of the other items is on the teacher. This Control dimension in the two-dimensional EFA is correlated 0.40 in the English sample and 0.42 in the math sample with the factor that is comprised of the remaining 29 items. Additionally, the standardized Chen and Thissen (1997) local dependence χ^2 indices

showed extremely large χ^2 values for the Control item set, indicating the existence of an important unmodeled dimension to explain the relationship between these items. In the three-dimensional EFA model, a small additional cluster appeared that contains the items B141, B29, B44, B89, and B135, which relate to enjoyment of learning in the classroom. This item cluster was found to be strongly correlated (.72 in English and 0.75 in the math sample) with the item cluster that contains the remaining 24 items.

I then performed a series of multilevel item factor analyses to test various hypothesized factor structures. Multilevel item factor analysis models consider hierarchically nested data wherein individuals are nested within a level-2 unit, such as classrooms. This is an improvement over traditional item factor analysis models, which assume that individuals are independent in the sample. These models can be thought of as confirmatory models, in that the researcher specifies the number of factors, the relations between factors and observed item responses a priori. In the multilevel item factor analysis model, the latent variables for individual i in classroom j are partitioned into two mutually exclusive parts: $\boldsymbol{\theta}_j = \left(\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}\right)$, where $\boldsymbol{\vartheta}_j$ is the vector of level-2 (classroom-level) latent variables and $\boldsymbol{\eta}_{ij}$ is the vector of individual-level (level-1) latent variables. The classroom-level latent factors represent "shared" perception within a class of teacher practices, and the student latent factors represent the students' latent deviation from the class section's shared perception of the teacher. The specification of the multilevel graded response model for ordinal item responses is provided as an appendix.

Based on the exploratory analyses and prior theory, four different models were compared. The first model is a multilevel unidimensional item factor analysis model in which all of the items load on a single dimension of teacher practice. The second model tested the Academic Support and Press framework that was described by Ferguson and Danielson (2014). This model specifies two correlated factors at each level of the model. The third model examines the structure suggested by the EFA models, which indicated the Control dimension is a separate (but correlated) dimension from the other Six Cs. Lastly, to examine the validity of the 7Cs structure, the final model that was estimated is a multilevel extension of the item bifactor model (Gibbons & Hedeker, 1992). In

this model, the common variance of each item is decomposed into contributions from a group-level general dimension (influencing all items in the domain), a within-level general dimension (influencing all items in the domain), and a group-specific dimension (influencing only the items within a cluster). The factors in the fourth model are uncorrelated, and due to estimation issues in this high-dimensional model, additional item parameter constraints were imposed on the specific factor slopes. Path diagrams for the four examined models are presented in Figure 5.1.

Model fit assessment for multilevel item factor analysis models is an open area of research, and many of the commonly-used methods that exist for single-level models are not currently available in existing software for multilevel item factor analysis models. In Table 5.5, the estimated models are compared using the -2log-likelihood, and two indices that are calculated using the log-likelihood: the Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978). AIC and BIC both incorporate penalties for model complexity, with the BIC imposing a stronger penalty. For all of the described model fit indices, lower values indicate the preferred model.

For both the math and English samples, parsimony (and the AIC and BIC) would suggest a preference for the unidimensional model. The next best fitting model is the Control and Six Cs correlated factors model.

As an additional check of the dimensionality of the Tripod survey, I used the standardized factor loading estimates from the multilevel bifactor model to calculate Explained Common Variance for a single Item (I-ECV; Stucky, Thissen, & Edelen, 2012). I-ECV describes the proportion of an item's common variance that is explained by the general dimension,

$$\text{I-ECV}_p = \frac{{\lambda_{pg}}^2}{{\lambda_{pg}}^2 + {\lambda_{ps}}^2}, \tag{5.1}$$

where λ_{pg} is the standardized factor loading for the pth item on the general dimension, λ_{ps} is the loading on the specific factor. In this scenario, items with higher I-ECVs indicate a stronger relationship to the overall teacher practice construct that is being measured by the general factor, and items with lower I-ECVs indicate that the item more strongly measures one of the seven specific dimension. Table 5.6 presents the

I-ECV values from the multilevel bifactor model for both subjects. For all of the items outside of the Control dimension, the I-ECV ranged from .87 to .99, indicating that these items are essentially unidimensional. The I-ECV for Control ranged from .12 to .78 in the English sample and .06 to .73 in Math, indicating these items are not strongly related to the overall dimension.

Given the EFA findings, as well as the multilevel item factor analysis and I-ECV results, the final model chosen is the correlated Six Cs and Control model. The Six Cs contain all of the items that fall within the Academic Support domain defined by Ferguson and Danielson (2014), as well as items that are intended to measure Challenge. Therefore, I refer from here on to the Six Cs composite as the Support dimension.

Class-level scores from Control and Support model are used in the subsequent validity steps. As a comparison, the classroom-level score based on student responses to all 36 items will also be analyzed in the subsequent validity steps, as this is the score that is typically used in summative teacher evaluations. Table 5.6 presents the standardized factor loadings for the unidimensional and Control and Six Cs confirmatory item factor analysis models for both subjects.

The second step in the Scoring inference is to examine bias in the classroom practice scores. A major concern with student surveys is bias resulting from how students use the response scale when rating teachers (Marsh, 1987). As described previously, (Popham, 2013) outlined sources of rater bias related to use of the response scale that could impact student ratings of teachers. If a teacher has a group of students that all follow a certain rating style, this may lead to systematic bias in the ratings of teachers.

This issue of idiosyncratic use of the response scale can be examined within the multilevel item factor analysis model. Traditional item factor analysis models assume that the item parameters (intercepts and slopes) are fixed coefficients (e.g., common to all respondents). By following the example of Maydeu-Olivares and Coffman (2006), I can add an additional student-level latent factor to explicitly model the students' tendency to use the response categories in a consistent but individually different manner. For example, the individual response idiosyncrasy latent factor can accommodate variance in responses due to students interpreting the category thresholds differently or

due to generosity error (the tendency to always use the high response categories when rating teachers). The item slopes for the random intercept η_{ij}^R are constrained equal for all items, and the variance of the latent factor is fixed equal to 1.0, which allows η_{ij}^R to represent individual differences in scale usage that are common across items. While the random intercept model is not a formal test of rater individual differences (as in Falk & Cai, 2015), it provides useful information about the degree to which student use the response scale differently.

In addition to simply accounting for the response scale differences, it is of interest to understand the student characteristics that are associated with idiosyncratic use of response scales. To model this association, I add a latent regression component to the model, allowing the random intercept latent variable η_{ij}^R to vary conditionally on a set of student background variables

$$\eta_{ij}^R = \mathbf{B}\mathbf{X}_{ij}^{'} + \epsilon_{ij}. \tag{5.2}$$

The multilevel item factor analysis model with the random intercept and latent regression is depicted graphically in Figure 5.2. The results of the latent regression are shown in Table 5.7. Of note, boys and gifted students are likely to use the response scale in a manner that is significantly different than female and non-gifted students. The direction and magnitude of the findings are similar for math and English classrooms.

The final step in the scoring inference is to examine scoring approaches. This paper compares three different scoring methods. Using the estimated item parameters from the multilevel item factor analysis calibration, estimates of the Control and Support scores are produced using Expected a Posteriori (EAP) scoring methods (Bock & Mislevy, 1982). For the jth classroom, let the EAP estimate for a given dimension be EAP(ϑ_j) and the corresponding standard errors be SE(ϑ_j). I compare the EAP scoring method to the class-level aggregate of the student scale scores and the adjusted mean score used in the MET study for the survey scales. The adjusted mean is produced

following a class-level regression

$$\overline{\text{SCALE}}_j = \beta_0 + \beta_1\overline{\text{TEST09}}_j + \beta_2\overline{\text{WHITE}}_j + \beta_3\overline{\text{SPED}}_j + \beta_4\overline{\text{FRPL}}_j + \beta_5\overline{\text{EL}}_j + \beta_6\overline{\text{GIFTED}}_j + \epsilon_j \quad (5.3)$$

where $\overline{\text{SCALE}}_j$ is the class-level simple mean of the student scale scores, which is regressed on classroom-level averages of student characteristics. The contextual variables that were included in this model are: class average prior year achievement ($\overline{\text{TEST09}}_j$), percentage students receiving special education services ($\overline{\text{SPED}}_j$), percentage of white students ($\overline{\text{WHITE}}_j$), percentage of students receiving free or reduced price lunch ($\overline{\text{FRPL}}_j$), percentage of English Learners ($\overline{\text{EL}}_j$), and percentage of gifted students ($\overline{\text{GIFTED}}_j$). The residuals e_j from this regression form the adjusted classroom level student perception survey scores.

These three scoring approaches were compared for three dimensions: (a) Control, (b) Support (the Six Cs), and (c) the Overall dimension. Classroom scores from each scoring approach were found to be strongly related to each other. The residual and sum score method scores are correlated between .94 and .98. The EAP scores are correlated with the other approaches between .85 and .93. For the subsequent validity results, EAP scores are used to represent the Control and Support dimensions, while the aggregated mean score is used for the Overall dimension. This decision was made because the Overall score (based on all the items) that is used in summative teacher evaluations is typically a simple mean score, so it is of interest to compare the reliability of this score as well as the model-based EAP scores.

5.6.2 Generalization

Multiple indices were used to estimate the reliability and stability of classroom practice scores, specifically focusing on three criteria (a) the interrelationship of items (Cronbach's α), (b) the reliability and within-group agreement of Tripod classroom scores (ρ_ϑ, ICC(1), and ICC(2)), and (c) correlations between scores across sections and over time (across-section and across-year correlation). Cronbach's α was estimated based on

student responses, ignoring the classroom structure of the data. A marginal reliability index ρ_ϑ summarizes the reliability of a measure as the proportion of variance in the observed score that is due to the true score

$$\rho_\vartheta = \frac{\sigma^2_\vartheta - \sigma^2_e(\vartheta)}{\sigma^2_\vartheta} = 1 - \frac{\sigma^2_e(\vartheta)}{\sigma^2_\vartheta} \tag{5.4}$$

where σ^2_ϑ is the prior value of the variance of ϑ, and $\sigma^2_e(\vartheta)$ is the marginal or average error variance of ϑ. The error variance $\sigma^2_e(\vartheta)$ can be computed from the estimated the standard errors in a sample of N respondents:

$$\overline{\hat{\sigma}}^2_e = \frac{1}{\mathrm{J}} \sum_{j=1}^{\mathrm{J}} \mathrm{SE}^2(\vartheta_j) \tag{5.5}$$

where $\mathrm{SE}^2(\vartheta_j)$ is the squared standard error for the jth classroom. Two indices are used to calculate group score reliability: ICC(1) and ICC(2). ICC(1) measures the proportion of the total variance that can be explained by group membership, and is calculated as $\frac{\tau_{00}}{\tau_{00}+\sigma^2}$, where τ_{00} represents between-group variance and σ^2 represents within-group variance. ICC(2), measuring the reliability of the group-mean, is calculated as (MSB-MSW)/MSB, where MSB is the between-group mean square from a one-way random effects ANOVA model, and MSW is the within-group mean square (Bliese, 2000). Lastly, for teachers that have multiple class sections in the MET sample, I estimate the proportion of variance in classroom scores that is "stable" by examining the correlation of scores within a teacher across class sections in a given year or between years.

Results from the Generalization inference are presented in Table 5.8. All of the dimensions were found to have high (>.80) Cronbach's α and marginal reliability. The ICC(1) values ranged from .27 to .39, indicating that the majority of variance in scores is within classrooms. However, the Control dimension had the highest ICC(1) estimates for both English and math, indicating that there is greater variance across teachers in this domain. The across-section correlation was over .50 for all dimensions, implying that more than half of the observed variation is due to stable differences between teachers. However, the across-year correlation in Tripod scores were low for

the English sample, ranging from .36-.51. Given that the Control dimensions contains items that are mostly focused on student behavior in the classroom rather than teacher characteristics, it is perhaps not surprising that this dimension has the consistently lowest across-section and across-year correlation.

Another concern within Generalization is the level of uncertainty in scores across the distribution of teacher effectiveness scores. Figures 5.3 and 5.4 display 95 percent confidence intervals for 100 randomly sampled classrooms from the full sample of English middle school classrooms for the Support and Control Tripod scores, respectively. The plots contain just a small sample of the total set of classrooms due to the difficulty of presenting close to 1,000 class sections in a single plot. In the figure, class sections are sorted by the Tripod score and the scores are plotted as circles. Ninety-five percent confidence intervals are plotted as horizontal lines for each class section. Red horizontal lines are plotted to show the 25th percentile and 75th percentile cut-offs. The width of the confidence intervals varies somewhat between classrooms, but generally stays consistent across the range of teacher effectiveness scores. These figures demonstrate that the confidence intervals for the classroom Tripod scores are mostly overlapping, with only a small subset of teachers that are definitively on the high or low end of the distribution. Comparing or ranking classrooms within the middle of the distribution would not be advised, as these score intervals are almost completely overlapping. Findings for math were very similar, and therefore are not shown here.

5.6.3 Extrapolation

First, I examined whether the Tripod scores are related to other measures of teacher effectiveness in the expected direction. Table 5.9 contains Pearson correlations of the Tripod scores with five other measures of teacher effectiveness: (a) domains from the Classroom Assessment Scoring System for secondary classrooms (CLASS), (b) domains from the Framework for Teaching (FFT), (c) class section averaged level of happiness and effort exerted in class, (d) class section value-added scores, and (e) a teacher-level Content Knowledge for Teaching score. Tripod scores were related to dimensions of CLASS and FFT in the expected direction. The Control dimension is more strongly

related to the CLASS classroom management domains (e.g., Behavior Management, Negative Climate) than the Support or Overall dimension, providing confirming evidence that Control is measuring classroom management behaviors. Not surprisingly, the Tripod dimensions are strongly related to the student-reported scales, which were collected as a part of the same student survey as the Tripod items. Interestingly, the Overall domain is very strongly related to happiness and effort exerted in class, while the Control dimension is only moderately related. All of the Tripod domains are significantly related to class section value-added model scores, with larger correlations seen in the math classrooms. There was no significant relationship between teacher Content Knowledge for Teaching scores and the Tripod scores.

Additionally, I examined the relationship between Tripod scores and student learning. Two different hierarchical linear model are specified in both English and math to measure the relationship between student achievement at the end of the 2009-10 school year (TEST10_{ij}) and the estimates of latent teacher effectiveness. In Model 1, Control and Support are included as classroom-level predictors of student achievement. In Model 2, the Overall Tripod score is included as a classroom-level predictor. The level-1 specification of Model 1 is summarized below

$$\begin{aligned} \text{TEST10}_{ij} =& \beta_{0j} + \beta_{1j}\text{TEST09}_{ij} + \beta_{2j}\text{WHITE}_{ij} + \beta_{3j}\text{FRPL}_{ij} + \\ & \beta_{4j}\text{GIFTED}_{ij} + e_{ij},\ e_{ij} \sim \text{N}\left(0, \sigma^2\right), \end{aligned} \tag{5.6}$$

where TEST10_{ij} is the estimate of student current end-of-year standardized (ELA/Math) test score and TEST09_{ij} is the estimate of student prior year standardized (ELA/Math) test score. Additionally, three student background characteristics are included: WHITE_{ij} is an indicator for whether the student is white, FRPL_{ij} is an indicator variable representing whether the student receives free or reduced price lunch (FPRL), and GIFTED_{ij} is an indicator for whether the student is labeled as gifted. Descriptive statistics for the student and classroom covariates are reported in Table 5.3. In the within-classroom model, all of the predictors are centered around group means so that β_{0j} represents the expected end-of-year standardized (English/Language Arts or Math) test score for class j. In addition, the slope coefficients for the level-1 predictors are treated as fixed

in my analyses.

At level-2, I model classroom-mean spring achievement β_{0j}:

$$\begin{aligned}\beta_{0j} =&\gamma_{00} + \gamma_{01}\overline{\text{TEST09}}_j + \gamma_{02}\overline{\text{SPED}}_j + \\ &+ \gamma_{03}\overline{\text{WHITE}}_j + \gamma_{04}\overline{\text{FRPL}}_j + \gamma_{05}\overline{\text{EL}}_j + \gamma_{06}\overline{\text{GIFTED}}_j \\ &+ \gamma_{07}\text{SUPPORT}_j + \gamma_{08}\text{CONTROL}_j + \mu_{0j},\ \mu_{0j} \sim N(0, \tau).\end{aligned} \tag{5.7}$$

At level 2, β_{0j} is modeled as a function of the grand mean, γ_{00}, class-level covariates, the Tripod domains (SUPPORT_j and CONTROL_j), and the random effect around the means, μ_{0j}. All of the predictors in Equation 5.6 are grand-mean centered. Unobserved differences between the school districts were controlled for by including district fixed effects in the model. The classroom-level residuals μ_{0j} is assumed to be normally distributed with means equal to zero and variance equal to τ. The key parameters of interest in the outcome model are γ_{07} and γ_{08}, which capture how the teacher practice dimensions relate to differences in class-mean end-of-year achievement, holding constant the other predictors in the model. The second model fit in each subject is the same as the first, except the Overall Tripod standardized mean score (OVERALL_j) is substituted into the model in place of the Tripod domains SUPPORT_j and CONTROL_j.

Table 5.10 presents the results from the two hierarchical linear models (Models 1 and 2) for English and Math. For both subjects, Control is significantly related to end-of-year student test scores. The Support dimension, comprised of the items from the other Six Cs, is not a significant predictor after controlling for the other variables in the model. That is to say, students who showed improved math and English achievement were in classrooms with higher reported levels of classroom management. In Model 2, the Overall factor was found to be significantly related to end-of-year student test scores for both subjects, but the coefficient for the Overall dimension is estimated to be substantially lower than for the Control dimension. These findings imply that the Control dimension should be of particular focus for school officials and policy makers who are trying to explain classroom differences in student growth.

5.6.4 Implication

This final step reviews whether the implications associated with the use of the Tripod scores to evaluate teacher effectiveness are appropriate. Scores are currently used in some school districts and states as a component of summative teacher evaluation decisions. The evidence collected in this study indicates that the Overall and Control scores show reasonable reliability and consistency across class sections, and that scores are related in expected direction to other measures of teaching quality. Given the confidence intervals in Figures 5.3 and 5.4, I do have evidence that most teachers' score confidence intervals are overlapping, so care should be taken if scores are going to be used to bin teachers into groups.

However, in the MET Project, no stakes are attached to scores, which limits the interpretations that can be made regarding the scores' summative purposes (e.g., identifying teachers to potentially remove or reward). Teachers in this study were volunteers who did not receive individual feedback from the study, and survey responses were collected following a procedure where students' responses could not be traced back to an individual. Changes to the data collection process for summative purposes might add additional concerns for the use of these scores for high-stakes decisions. Districts could collect additional information regarding the fairness and accuracy of the data collection procedures related to summative uses of the Tripod survey, focusing on issues of confidentiality (are students told that responses are confidential?), accuracy of attribution (has the student rating the teacher actually been in the classroom long enough to fairly evaluate?), and bias (is the teacher in the classroom during the survey, and if so, does he or she provide any additional instructions that may bias results?). These additional pieces of evidence would help bolster the validity argument for using Tripod scores to sort teachers for summative purposes.

Tripod scores are also being used to provide feedback for professional development purposes. Again, the MET study was not designed to allow for direct investigation of the effectiveness of Tripod score use for feedback and coaching. The evidence needed to support this use could focus on two questions: (a) are scores provided to teachers in a way that allows for the understanding of strengths and weaknesses, and (b) are

coaching systems in place to support teacher growth? A report by the Bill & Melinda Gates Foundation (2012) provides examples of efforts by the Green Dot Public schools to meaningfully present Tripod results. The degree to which results are meaningful will depend on the score reports that are used within the district or school. If the Tripod is to be used to route teachers into a professional development program and to monitor the teacher growth, additional evidence would be needed to demonstrate the Tripod's sensitivity to measuring growth in teacher effectiveness.

5.7 Conclusions

This article applies the validity argument approach outlined by Kane (2006) to the use of the Tripod secondary survey to measure teacher effectiveness. Data from the Measures of Effective Teaching (MET) Project was used to establish evidence of dimensionality, bias, reliability, and concurrent validity of the Tripod scores.

The validity argument consists of four steps. The scoring argument holds that teacher scores produced from the Tripod survey accurately capture dimensions of instructional quality undistorted by other factors. I found that the 7Cs theoretical structure is not well supported. However, the Control (classroom management) dimension was found to be a separate factor. Classroom management also appears as a central domain in the CLASS classroom observational protocol, where it is referred to as "Classroom Organization" (sub-domains include behavior management, productivity, instructional learning formats), and the Framework for Teaching, where it is called "Classroom Environment" (sub-domains include managing classroom procedures and managing student behavior) (Kane & Staiger, 2012). The importance of classroom management as a separate dimension that is predictive of student learning is supported by a body literature (Evertson, Emmer, Sanford, & Clements, 1983; Emmer & Stough, 2001). The other Six Cs (Care, Clarify, Challenge, Captivate, Confer and Consolidate) can be grouped together as a single measure of Support. The Support dimension contains items that measure multiple types of teacher support that have been previously been found to be important in the teacher quality literature, including instructional support, and socio-emotional support (Lee, Smith, Perry, & Smylie, 1999;

Pianta & Hamre, 2009). Given that six of the 7Cs are found to form a single composite measure of teacher support, it is likely that this domain could be reliably measured in less than the current 29 item length. Approaches to produce reliable short forms will be investigated in future studies.

The second step is concerned with the generalization of scores that might have been obtained with a different group of students or a different set of items. The reliability/stability of the Tripod far exceeds that of value-added scores (Kane & Cantrell, 2010). The average across-section correlation for the Tripod Support scores was 0.67 for both English and Math, and the across-section correlation for the Control was 0.53 for English and 0.60 for Math. These findings indicate that the reliability of one administration of the Tripod survey far exceeds a single classroom observation rating. Estimation of reliability for classroom observation protocols depend on multiple criteria, including the number of raters and number of lessons scored. Using highly-trained raters who rated classroom observation videos, the MET study found that for most of the observational protocols, at least four rated lessons, each with a different rater, were required to achieve reliability in the neighborhood of 0.65 (Kane & Staiger, 2012).

However, using the MET data, we are unable to look at the stability of ratings within a school year. All of the student ratings in this study were provided in the spring of the school year. Ferguson (2010) has previously reported that the correlation between student ratings across two time points (March and December) within a school year ranged between 0.70 and 0.85. Therefore, we expect there to be some non-zero variance due to rating occasion. A data collection design with student ratings across multiple occasions and multiple section within a school year would allow for a more thorough discussion of the generalizability of scores. This design would provide more information about the percentage of the variance in the teacher scores that is attributable to persistent differences among teachers.

In the third step, I examined the extrapolation of the Tripod survey scores to a more broadly construed concept of teacher effectiveness. The Tripod Control scores were found to be significant predictors of student academic achievement in both English and Math, though Support was not a significant predictor. Significant correlations

in the expected directions were also found between the Tripod Overall, Control, and Support dimensions and dimensions of both the Classroom Assessment Scoring System (CLASS) and Framework for Teaching observational protocols.

Given all of these findings, I have preliminary evidence to suggest that the Tripod survey provides a reliable measure of teacher support and classroom management. Additional evidence will need to be collected to determine whether it is reliable enough support the use of the Tripod for high-stakes purposes. It is clear from the current findings that when communicating scores to teachers and stake-holders, standard errors of measurement should be included in addition to the scores to provide context for comparisons across teachers. The caterpillar plots shown in Figures 5.3 and 5.4 make it clear that rank-ordering or binning teachers who fall in the middle of the distribution will lead to errors in judgment due to the uncertainty in scores.

Additionally, this study provides insights on uses of the Tripod in a professional development context. Though the theoretical model of the Tripod 7Cs may still be valid, I did not find that the 7Cs could be reliably differentiated with middle school students' ratings of their classrooms. Therefore, it is recommended that scores are not reported separately for Care, Clarify, Challenge, Captivate, Confer and Consolidate, as these scores are highly correlated and each less reliable than the composite Support score. Furthermore, if Tripod scores are used to measure growth over time in response to professional development, confidence intervals around scores provide context for whether any observed growth falls within the range of uncertainty in scores.

The validity argument approach applied to the Tripod survey requires clearly outlining the warrants and claims for the instrument's validity. This is an important step for clarify the purpose and acceptable users of the measure. However, there are limitations to this approach when examining the validity of an instrument using large data sets such as the Measures of Effective Teaching data set. Various aspects of the survey administration, including whether it is administered on paper or online, whether the teacher is nearby as students fill out the survey, whether it is administered near the start of the year or end of the year, could all affect the validity and reliability of the measure. However, this kind of administration information is not generally available

in large datasets. Additionally, large studies such as the MET typically collect data for research purposes, so the conditions of data collection may be far different than a typical operational use of the survey.

This study lays the groundwork for a validity argument approach for the Tripod survey. Though the present investigation cannot demonstrate the validity of this survey for high-stakes purposes, school districts that are collecting the Tripod or another student survey measure as a component of a summative teacher evaluation could follow the validity argument approach to illustrate the degree to which the reliability and validity conclusions change when stakes are added.

Table 5.1: An interpretive argument for Tripod scores

Stage of argument	Description	Sub-components of the stage	Focusing questions
Scoring	The intent of this stage is to demonstrate that the Tripod dimensions measure the teacher's true instructional quality	Dimensionality assessment	What dimensions of teacher quality can be captured by the Tripod survey?
		Examination of bias	Is there bias in items due to observed group characteristics?
		Scoring	Are the scores unbiased (e.g., is systematic error small?) Are teachers in certain settings less likely to receive high scores than other teachers due to outside factors?
Generalization	Generalization addresses test reliability and stability	IRT-based reliability	How reliable are inferences across the range of the latent trait?
		Stability of teacher scores across time	Are a teacher's Tripod survey scores correlated across years?
Extrapolation	Extrapolating to a more broad definition of teacher effectiveness	Correlational analyses with VAM and teacher observation scores	How strongly due student-reported measures of teacher effectiveness correlated with other teacher quality measures?
		Hierarchical linear models predicting student academic gains by Tripod teacher scores	Which dimensions of teacher quality are most predictive of student academic growth?
Implication	Soundness of intended interpretation of scores	n/a	What is the rationale for using Tripod scores to evaluate teachers? Are there unintended consequences of using these scores?

Table 5.2: Tripod 7Cs and description of the domains

Dimension	Description
Control	Control concerns the degree to which the class is both well-behaved and on task.
Care	Care concerns whether teachers have supportive relationships with students.
Clarify	Clarify concerns how effectively the teacher is able to help students understand concepts taught in class.
Challenge	Challenge concerns both effort and teacher's insistence that students persist in the face of difficulty.
Captivate	Captivate pertains to how effectively the teacher is able to hold the students' attention in class.
Confer	Confer concerns the level to which a teacher both gets students to provide their ideas and welcomes their feedback.
Consolidate	Consolidate concerns making learning coherent, giving feedback, and checking for understanding.

Table 5.3: Student, classroom, and teacher demographic variables

	English Classrooms				Math Classrooms			
	2009–10		2010–11		2009–10		2010–11	
Category Name	M	SD	M	SD	M	SD	M	SD
Student Characteristics	N=19,406		N=8,153		N=16,716		N=7,710	
Male	0.13	0.34	0.14	0.34	0.50	0.50	0.50	0.50
Gifted	0.13	0.34	0.15	0.36	0.12	0.32	0.11	0.31
English Learner (EL)	0.06	0.24	0.06	0.24	0.14	0.35	0.13	0.34
White	0.26	0.44	0.27	0.44	0.27	0.44	0.26	0.44
Black	0.28	0.45	0.28	0.45	0.29	0.45	0.27	0.44
Hispanic	0.36	0.48	0.34	0.47	0.35	0.48	0.36	0.48
Free/reduced lunch	0.62	0.49	0.62	0.49	0.61	0.49	0.61	0.49
Special Education	0.49	0.50	0.49	0.50	0.07	0.26	0.07	0.25
Prior year test score	0.16	0.94	0.21	0.93	0.13	0.94	0.20	0.89
End of year test score	0.17	0.94	0.19	0.93	0.15	0.94	0.16	0.91
Classroom Characteristics	N=1,071		N=406		N=946		N=380	
% Gifted	0.12	0.21	0.12	0.22	0.10	0.20	0.10	0.19
% ELL	0.14	0.18	0.15	0.19	0.15	0.17	0.15	0.18
% White	0.25	0.28	0.25	0.28	0.25	0.28	0.24	0.27
% Black	0.30	0.29	0.29	0.29	0.31	0.28	0.29	0.29
% Hispanic	0.37	0.31	0.37	0.30	0.37	0.29	0.37	0.28
% FRPL	0.64	0.30	0.65	0.29	0.65	0.30	0.64	0.29
Prior year test score	0.07	0.13	0.07	0.13	0.09	0.14	0.08	0.13
Number of students	23.6	6.10	24.9	6.50	23.8	7.10	25.6	7.20
Teacher Characteristics	N=572		N=406		N=504		N=380	
Years of experience	8.08	7.26	7.93	7.18	7.49	7.18	7.81	7.20
Male	0.15	0.36	0.16	0.37	0.29	0.45	0.26	0.44
White	0.58	0.49	0.56	0.50	0.51	0.50	0.54	0.50
Black	0.36	0.48	0.37	0.48	0.37	0.48	0.35	0.48
Hispanic	0.05	0.22	0.05	0.22	0.06	0.24	0.06	0.24
Master's degree or higher	0.30	0.46	0.28	0.45	0.28	0.45	0.31	0.46
Value-added score	0.00	0.14	0.01	0.14	0.01	0.20	0.02	0.20

Table 5.4: Item wording and descriptive statistics for the 36 Tripod student survey items

Domain	Item	Item wording	ELA		Math	
			M	SD	M	SD
Care	A10	My teacher in this class makes me feel that s/he really cares about me.	3.72	1.22	3.64	1.25
Care	B146	My teacher seems to know if something is bothering me.	3.14	1.31	3.05	1.32
Care	B34	My teacher really tries to understand how students feel about things.	3.60	1.18	3.47	1.22
Control	B112	Student behavior in this class is under control.	3.39	1.23	3.36	1.26
Control	B113*	I hate the way that students behave in this class.	3.53	1.32	3.49	1.34
Control	B114*	Student behavior in this class makes the teacher angry.	3.00	1.31	2.92	1.33
Control	B138*	Student behavior in this class is a problem.	3.34	1.26	3.31	1.29
Control	B46	My classmates behave the way my teacher wants them to.	3.15	1.19	3.10	1.22
Control	B49	Students in this class treat the teacher with respect.	3.62	1.13	3.55	1.17
Control	B6	Our class stays busy and does not waste time.	3.48	1.13	3.50	1.17
Clarify	B1	If you don't understand something, my teacher explains it another way.	4.04	1.04	4.03	1.08
Clarify	B130	My teacher knows when the class understands, and when we do not.	3.81	1.07	3.82	1.11
Clarify	B136	When s/he is teaching us, my teacher thinks we understand even when we don't.	3.62	1.19	3.54	1.23
Clarify	B17	My teacher has several good ways to explain each topic that we cover in this class.	3.94	1.04	3.91	1.10
Clarify	B80	My teacher explains difficult things clearly.	3.88	1.07	3.85	1.13
Challenge	B128	My teacher asks questions to be sure we are following along when s/he is teaching.	4.30	0.96	4.39	0.94
Challenge	B133	My teacher asks students to explain more about answers they give.	4.07	0.97	4.12	0.97
Challenge	B21	In this class, my teacher accepts nothing less than our full effort.	4.01	1.05	4.02	1.06

Challenge	B36	My teacher doesn't let people give up when the work gets hard.	4.02	1.09	4.03	1.09
Challenge	B45	My teacher wants us to use our thinking skills, not just memorize things.	4.13	0.99	4.10	1.02
Challenge	B59	My teacher wants me to explain my answers–why I think what I think.	4.10	1.00	4.07	1.03
Challenge	B70	In this class, we learn a lot almost every day.	3.81	1.07	3.99	1.04
Challenge	B90	In this class, we learn to correct our mistakes.	4.05	1.02	4.08	1.02
Captivate	B141*	This class does not keep my attention–I get bored.	3.42	1.31	3.37	1.35
Captivate	B29	My teacher makes learning enjoyable.	3.62	1.23	3.49	1.29
Captivate	B44	My teacher makes lessons interesting.	3.61	1.21	3.48	1.26
Captivate	B89	I like the ways we learn in this class.	3.83	1.00	3.83	1.03
Confer	B129	My teacher wants us to share our thoughts.	3.95	1.10	3.68	1.19
Confer	B135	Students get to decide how activities are done in this class.	2.45	1.07	2.30	1.07
Confer	B154	My teacher gives us time to explain our ideas.	3.78	1.08	3.66	1.13
Confer	B155	Students speak up and share their ideas about class work.	3.64	1.15	3.54	1.18
Confer	A54	My teacher respects my ideas and suggestions.	3.83	1.12	3.71	1.15
Consol.	B145	My teacher takes the time to summarize what we learn each day.	3.49	1.19	3.50	1.23
Consol.	B147	My teacher checks to make sure we understand what s/he is teaching us.	4.04	1.05	4.09	1.06
Consol.	B58	We get helpful comments to let us know what we did wrong on assignments.	3.81	1.14	3.69	1.19
Consol.	B83	The comments that I get on my work in this class help me understand how to improve.	3.84	1.11	3.69	1.16

Note. "Consol." is an abbreviation for Consolidate.

* signifies an item that has been reverse-coded.

Table 5.5: Comparison of the unidimensional, Press & Support, Control & Support (Six Cs), and bifactor measurement models

Subject	Model	-2loglikelihood	AIC	BIC
English	Unidimensional	1658555, 1658633	1658917, 1658995	1660339, 1660416
	Press and Support	1792033, 1792055	1792395, 1792417	1793817, 1793838
	Control and Support (Six Cs)	1698934, 1698963	1699302, 1699331	1700748, 1700776
	Bifactor model with 7Cs	1864007, 1864075	1864455, 1864523	1866213, 1866281
Math	Unidimensional	1476059, 1476074	1476421, 1476436	1477819, 1477834
	Press and Support	1585091, 1585112	1585453, 1585474	1586852, 1586872
	Control and Support (Six Cs)	1500456, 1500481	1500824, 1500849	1502245, 1502271
	Bifactor model with 7Cs	1646639, 1646715	1647087, 1647163	1648817, 1648893

Note. The -2log-likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) for the multilevel item factor models are reported as 95 percent confidence intervals. Using the MH-RM algorithm, the confidence intervals are approximated based on 1000 Monte Carlo draws. Given the reported values represent an approximation of the log-likelihood, caution should be used in comparing estimates across models.

Table 5.6: Multilevel item Factor analysis results with item explained common variance (I-ECV)

Tripod Domain	Item	English				Math			
		Unid.	6Cs & Control		I-	Unid.	6Cs & Control		I-
		λ_1	λ_1	λ_2	ECV	λ_1	λ_1	λ_2	ECV
Care	A10	.67(.01)	.68(.01)	0(—)	.97	.66(.01)	.67(.01)	0(—)	.98
Care	B146	.57(.01)	.58(.01)	0(—)	.95	.56(.01)	.57(.01)	0(—)	.95
Care	B34	.69(.01)	.70(.01)	0(—)	.98	.66(.01)	.67(.01)	0(—)	.98
Control	B112	.43(.01)	0(—)	.60(.01)	.69	.41(.01)	0(—)	.59(.01)	.62
Control	B113	.17(.01)	0(—)	.46(.01)	.12	.15(.01)	0(—)	.45(.01)	.06
Control	B114	.31(.01)	0(—)	.52(.01)	.46	.27(.01)	0(—)	.50(.01)	.33
Control	B138	.30(.01)	0(—)	.57(.01)	.43	.28(.01)	0(—)	.57(.01)	.33
Control	B46	.49(.01)	0(—)	.64(.01)	.76	.46(.01)	0(—)	.63(.01)	.69
Control	B49	.51(.01)	0(—)	.63(.01)	.78	.49(.01)	0(—)	.62(.01)	.73
Control	B6	.48(.01)	0(—)	.49(.01)	.75	.44(.01)	0(—)	.48(.01)	.67
Clarify	B1	.67(.01)	.68(.01)	0(—)	.99	.65(.01)	.66(.01)	0(—)	.99
Clarify	B130	.60(.01)	.61(.01)	0(—)	.99	.59(.01)	.59(.01)	0(—)	.98
Clarify	B136	.40(.01)	.40(.01)	0(—)	.96	.41(.01)	.41(.01)	0(—)	.94
Clarify	B17	.70(.01)	.71(.01)	0(—)	.99	.69(.01)	.70(.01)	0(—)	.99
Clarify	B80	.69(.01)	.70(.01)	0(—)	.99	.69(.01)	.70(.01)	0(—)	.99
Chall.	B128	.55(.01)	.56(.01)	0(—)	.97	.53(.01)	.54(.01)	0(—)	.96
Chall.	B133	.49(.01)	.49(.01)	0(—)	.95	.46(.01)	.47(.01)	0(—)	.95
Chall.	B21	.55(.01)	.55(.01)	0(—)	.97	.54(.01)	.54(.01)	0(—)	.96
Chall.	B36	.62(.01)	.63(.01)	0(—)	.98	.61(.01)	.61(.01)	0(—)	.98
Chall.	B45	.59(.01)	.6(.01)	0(—)	.98	.56(.01)	.57(.01)	0(—)	.97
Chall.	B59	.57(.01)	.58(.01)	0(—)	.97	.53(.01)	.54(.01)	0(—)	.96
Chall.	B70	.64(.01)	.65(.01)	0(—)	.98	.60(.01)	.60(.01)	0(—)	.97
Chall.	B90	.67(.01)	.68(.01)	0(—)	.99	.66(.01)	.66(.01)	0(—)	.98
Capt.	B141	.53(.01)	.53(.01)	0(—)	.89	.52(.01)	.52(.01)	0(—)	.89
Capt.	B29	.70(.01)	.71(.01)	0(—)	.97	.70(.01)	.71(.01)	0(—)	.97
Capt.	B44	.71(.01)	.72(.01)	0(—)	.97	.71(.01)	.71(.01)	0(—)	.97
Capt.	B89	.64(.01)	.64(.01)	0(—)	.95	.64(.01)	.64(.01)	0(—)	.95
Confer	B129	.51(.01)	.52(.01)	0(—)	.94	.49(.01)	.50(.01)	0(—)	.94
Confer	B135	.38(.01)	.38(.01)	0(—)	.87	.39(.01)	.40(.01)	0(—)	.88
Confer	B154	.68(.01)	.69(.01)	0(—)	.98	.66(.01)	.67(.01)	0(—)	.98
Confer	B155	.58(.01)	.58(.01)	0(—)	.96	.56(.01)	.57(.01)	0(—)	.96
Confer	A54	.69(.01)	.70(.01)	0(—)	.98	.68(.01)	.68(.01)	0(—)	.98
Consol.	B145	.61(.01)	.62(.01)	0(—)	.99	.59(.01)	.60(.01)	0(—)	.98
Consol.	B147	.72(.01)	.73(.01)	0(—)	.99	.70(.01)	.70(.01)	0(—)	.99
Consol.	B58	.64(.01)	.65(.01)	0(—)	.99	.63(.01)	.64(.01)	0(—)	.99
Consol.	B83	.65(.01)	.66(.01)	0(—)	.99	.64(.01)	.65(.01)	0(—)	.99

Note. Only between-level standardized factor loadings are reported in this table. Due to cross-level invariance constraints, the within-level loadings are equal to the values in the table.

Table 5.7: Latent regression results from the multilevel item factor analysis models examining response bias

Variable	English	Math
Age	.01(.00)	.01(.00)
Male	-.27(.03)	-.23(.02)
Gifted	.50(.08)	.40(.03)
Special education	.01(.05)	.06(.02)
EL	-.20(.05)	-.12(.04)
FRPL	-.03(.05)	-.01(.00)
Black	.04(.05)	.10(.03)
Hispanic	-.06(.04)	-.12(.03)

Table 5.8: Reliability statistics for the Control, Support, and Overall Tripod scores

	ELA			Math		
	Support Score[a]	Control Score[a]	Overall Score[b]	Support Score[a]	Control Score[a]	Overall Score[b]
Cronbach's Alpha	0.95	0.84	0.84	0.85	0.85	0.95
Marginal Reliability	0.85	0.90	0.87	0.86	0.91	0.85
ICC(1)	0.27	0.35	0.30	0.28	0.39	0.32
ICC(2)	0.84	0.90	0.86	0.85	0.91	0.87
Across Section Correlation (Year 1 only)	0.67	0.53	0.63	0.67	0.6	0.62
Across-Year Correlation	0.51	0.36	0.49	0.59	0.59	0.57

[a] The Support and Control scores are calculated as posterior means from a multilevel item factor model based on the factor loadings reported in Table 5.6.

[b] The overall score is calculated as an aggregated sum score, first averaging student responses for all 36 items, and then averaging student scores to the classroom level. This is the score typically reported in teacher evaluation systems.

Table 5.9: Pearson correlations between Tripod domain scores and other measures of teaching quality

Instrument and domain (if applicable)	ELA			Math		
	Support Score[a]	Control Score[a]	Overall Score[b]	Support Score[a]	Control Score[a]	Overall Score[b]
CLASS						
Positive climate	0.29**	0.20**	0.30**	0.28**	0.19**	0.26**
Negative climate	-0.25**	-0.37**	-0.33**	-0.12**	-0.35**	-0.16**
Teacher sensitivity	0.24**	0.16**	0.29**	0.24**	0.09*	0.22**
Regard for student thoughts	0.20**	0.10*	0.27**	0.26**	0.09	0.24**
Behavior management	0.25**	0.41**	0.33**	0.16**	0.39**	0.23**
Productivity	0.21**	0.26**	0.25**	0.17**	0.24**	0.19**
Instructional learning formats	0.22**	0.15**	0.30**	0.21**	0.09*	0.19**
Content understanding	0.20**	0.16**	0.27**	0.14**	0.08	0.11*
Analysis and problem solving	0.18**	0.12**	0.26**	0.18**	0.04	0.15**
Quality of feedback	0.27**	0.12*	0.31**	0.24**	0.09*	0.21**
Instructional dialogue	0.26**	0.10*	0.28**	0.24**	0.03	0.18**
Student Engagement	0.32**	0.20**	0.32**	0.29**	0.23**	0.27**
Framework for Teaching						
Creating respect	0.12**	0.31**	0.21**	0.17**	0.31**	0.22**
Using questioning techniques	0.14**	0.13**	0.15**	0.12**	0.18**	0.18**
Establishing learning culture	0.10*	0.24**	0.20**	0.18**	0.28**	0.24**
Managing class procedures	0.09*	0.25**	0.16**	0.08	0.27**	0.14**
Communicating with students	0.13**	0.23**	0.17**	0.15**	0.18**	0.20**
Managing student behavior	0.13**	0.36**	0.23**	0.12**	0.40**	0.20**
Engaging students	0.16**	0.28**	0.21**	0.18**	0.20**	0.21**
Using tests in instruction	0.15**	0.20**	0.19**	0.15**	0.18**	0.23**
Student-reported scales						
Happiness in class	0.67**	0.39**	0.82**	0.67**	0.40**	0.64**
Effort exerted in class	0.67**	0.46**	0.67**	0.78**	0.42**	0.80**
Other						
VAM score	0.09**	0.16**	0.13**	0.18**	0.25**	0.27**
Teaching Content Knowledge	0.07	0.02	0.04	0.01	0.05	0.06

Note. CLASS-S=Classroom Assessment Scoring System. * Correlation is significant at the .05 level, two-tailed. ** Correlation is significant at the .01 level, two-tailed.

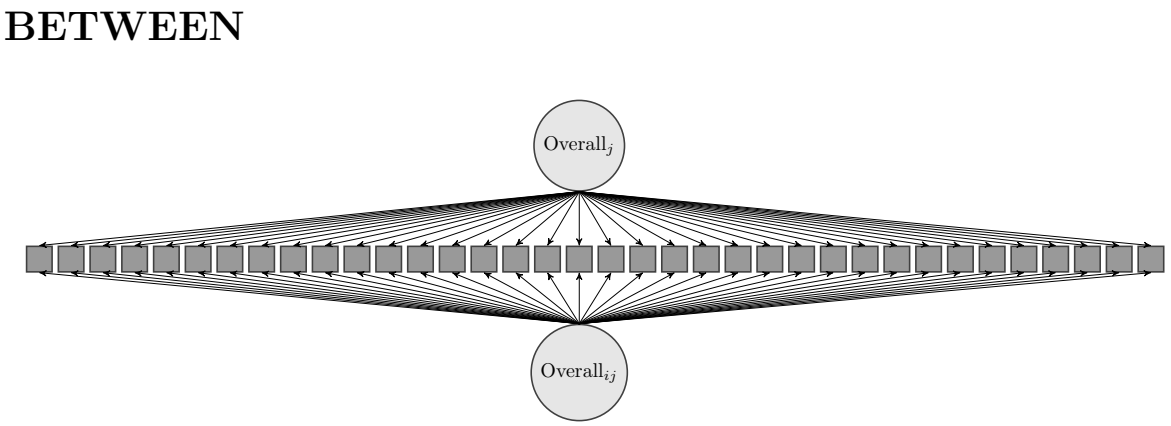
Table 5.10: Results from the multilevel model predicting end-of-year (spring 2010) student achievement

	ELA		Math	
	Model 1	Model 2	Model 1	Model 2
(Intercept)	0.04(0.02)*	0.04(0.02)**	0.02(0.02)	0.02(0.02)
Prior year test score (spring 2009)	0.64(0.01)***	0.64(0.01)***	0.68(0.01)***	0.68(0.01)***
Gifted	0.21(0.02)***	0.21(0.02)***	0.18(0.02)***	0.19(0.02)***
White	0.05(0.02)***	0.05(0.02)***	0.03(0.02)*	0.03(0.02)*
Free or Reduced Price Lunch (FRPL)	-0.06(0.01)***	-0.06(0.01)***	-0.04(0.01)**	-0.04(0.01)***
Class-level average prior year test score	0.61(0.02)***	0.61(0.02)***	0.66(0.02)***	0.67(0.02)***
% Gifted	0.24(0.04)***	0.24(0.04)***	0.14(0.05)**	0.14(0.05)**
% White	0.06(0.04)	0.07(0.04)	0.04(0.05)	0.04(0.04)
% Free or Reduced Price Lunch (FRPL)	-0.03(0.04)	-0.03(0.04)	-0.06(0.04)	-0.07(0.04)
% Special Education	-0.11(0.07)	-0.10(0.07)	-0.04(0.06)	-0.06(0.07)
% English Learner	0.03(0.05)	0.04(0.05)	0.07(0.05)	0.07(0.05)
Support	-0.04(0.03)	—(—)	0.00(0.03)	—(—)
Control	0.14(0.02)***	—(—)	0.18(0.02)***	—(—)
Overall	—(—)	0.04(0.01)***	—(—)	0.07(0.01)***

Note. N (English classrooms) = 12,562. N (Math classrooms) = 9,496.
*p<.05. **p<.01. ***p<.001

Figure 5.1: Path diagrams for the four confirmatory multilevel item factor analysis models

(a) Unidimensional model



(b) Press and Support model

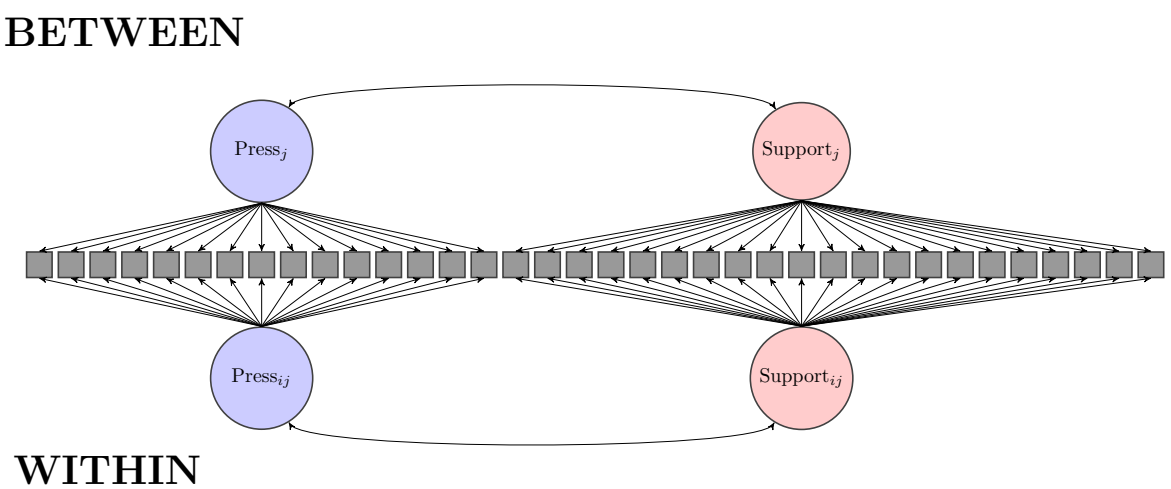
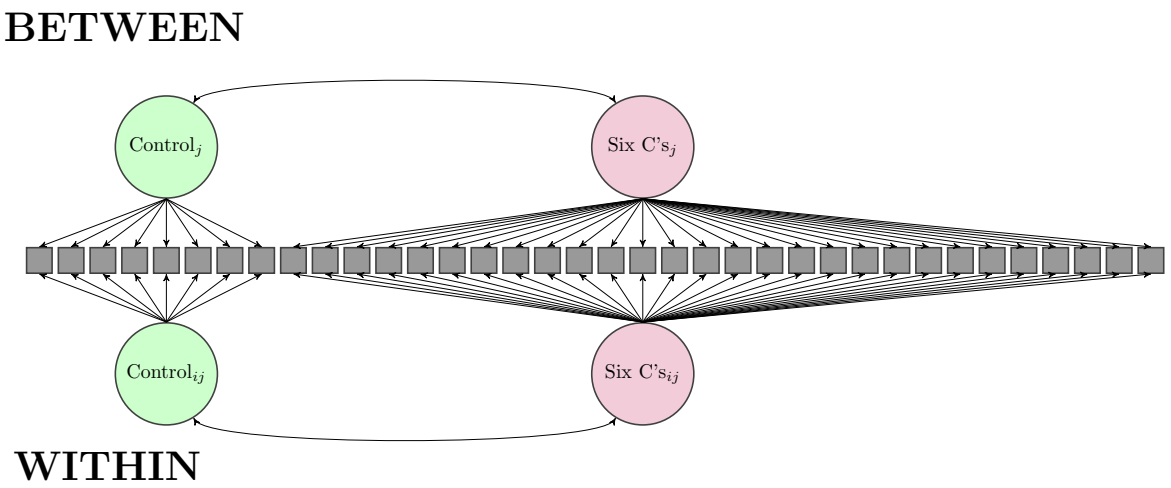


Figure 5.1: Path diagrams for the four confirmatory multilevel item factor analysis models–Continued

(c) Control and Support (Six Cs) model



(d) Bifactor model with the 7Cs structure

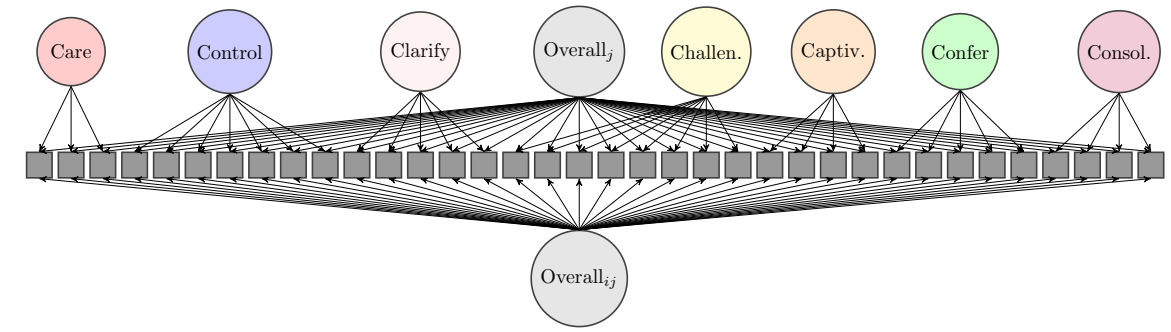


Figure 5.2: Path diagram for the multilevel item factor analysis model with random intercept

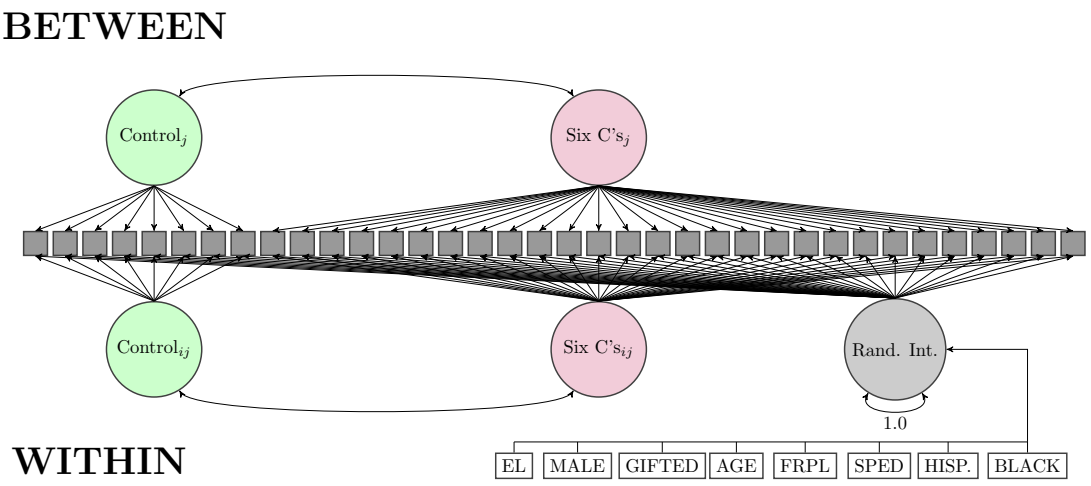
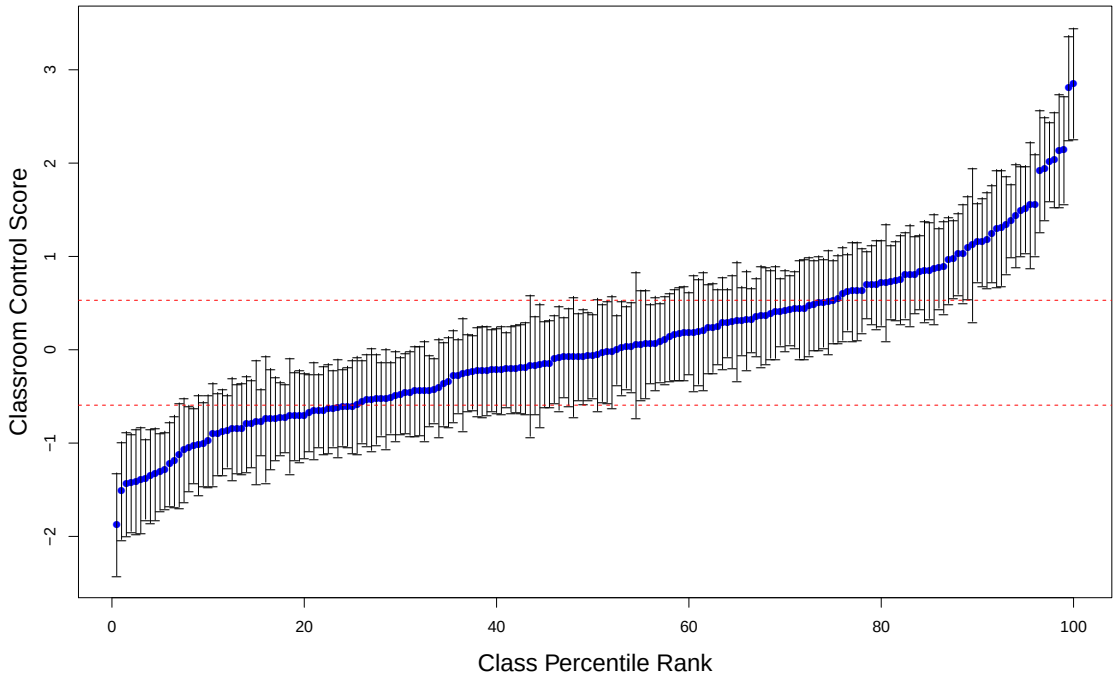
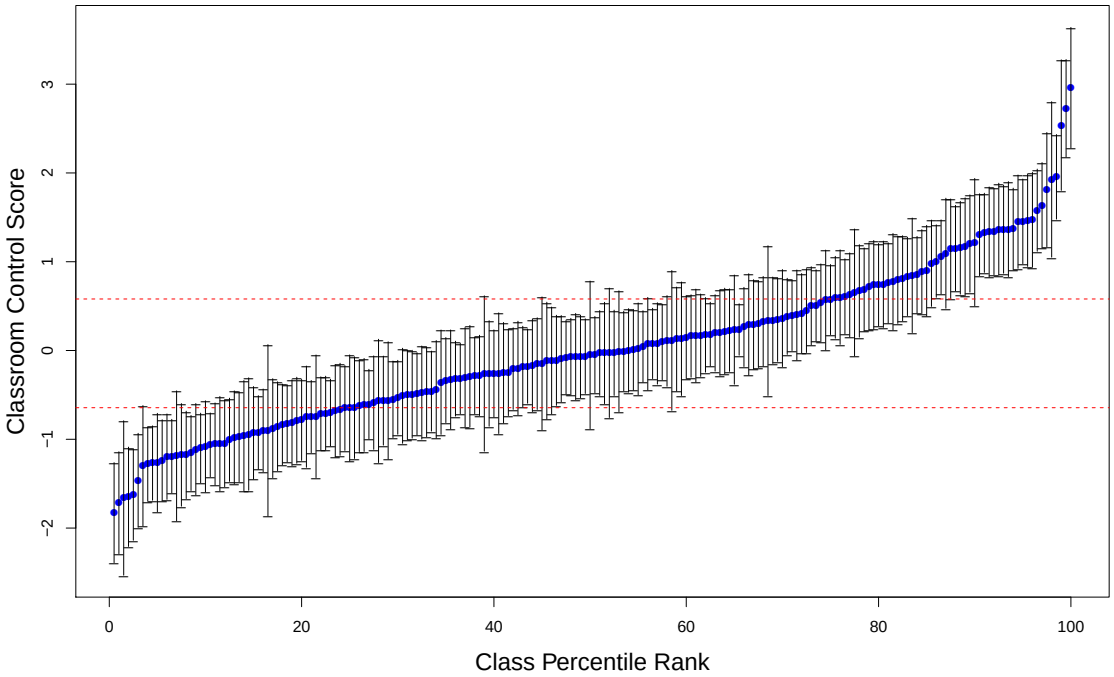


Figure 5.3: Confidence intervals for the Support Tripod scores for 100 sampled classrooms from the English sample



NOTE: The blue dot represents the class section estimated score, and the black line around the score represents the 95 percent confidence interval around the score. The horizontal red dashed lines represent the 25th and 75th percentile scores.

Figure 5.4: Confidence intervals for the Support Tripod scores for 100 sampled classrooms from the math sample.



NOTE: The blue dot represents the class section estimated score, and the black line around the score represents the 95 percent confidence interval around the score. The horizontal red dashed lines represent the 25th and 75th percentile scores.

5.8 Appendix

Let there be $p = 1, \ldots, n$ items. A student i is nested in a group j. Let the response from individual i in group j to item p be y_{pij}, where y_{pij} has K_p response categories, so that $y_{pij} \in (0, \ldots, K_p - 1)$. Let there be $i = 1, ..., N_j$ students in group j, with $j = 1, \ldots, J$ groups. The overall sample size is $N = \sum_{j=1}^{J} N_j$.

In this model, the latent variables for individual i in classroom j are partitioned into two mutually exclusive parts: $\boldsymbol{\theta}_{ij} = \left(\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}\right)$, where $\boldsymbol{\vartheta}_j$ is the vector of group-level (level-2) latent variables, and $\boldsymbol{\eta}_{ij}$ is the vector of individual-level (level-1) latent variables. The multilevel item factor (IFA) model specifies the conditional probability for the response to item p with K_p categories from student i in classroom j. I use a multidimensional extension of the graded response model (Samejima, 1969). The category response probability is the difference between adjacent cumulative probabilities

$$\begin{aligned} P(y_{pij} \geq 1|\boldsymbol{\vartheta}_j, \eta_{ij}) &= \frac{1}{1 + \exp\left[-\left(c_{p,1} + \mathbf{a}_p^{B'}\boldsymbol{\vartheta}_j + \mathbf{a}_p^{W'}\eta_{ij}\right)\right]} \\ &\vdots \\ P(y_{pij} \geq K - 1|\boldsymbol{\vartheta}_j, \eta_{ij}) &= \frac{1}{1 + \exp\left[-\left(c_{p,K-1} + \mathbf{a}_p^{B'}\boldsymbol{\vartheta}_j + \mathbf{a}_p^{W'}\eta_{ij}\right)\right]} \end{aligned} \tag{5.8}$$

The item parameters for item p include: a set of K-1 (strictly ordered) intercepts $c_{p,1}, \ldots, c_{p,K-1}$, a conformable vector of level-1 slopes $\mathbf{a}_p^{B'}$, and a conformable vector of level-2 item slopes $\mathbf{a}_p^{W'}$. The slopes (or discrimination) parameters are analogous to item factor loadings. The category response probability is the difference between two adjacent cumulative probabilities

$$P(y_{pij} = k|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}) = P(y_{pij} \geq k|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}) - P(y_{pij} \geq k + 1|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij}) \tag{5.9}$$

where $P(y_{pij} \geq 0|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij})$ is equal to 1 and $P(y_{pij} \geq K_p|\boldsymbol{\vartheta}_j, \boldsymbol{\eta}_{ij})$ is zero.

Given $p = 1, ..., n$ items, $i = 1, ..., N_j$ individuals in classroom j, and $j = 1, \ldots, J$ classrooms, the observed data (marginal) likelihood function may take the following

form:

$$L\left(\boldsymbol{\gamma}\right)=\prod_{j=1}^{J}\int\prod_{i=1}^{N_j}\left[\int\prod_{p=1}^{n}P(Y_{pij}=y_{pij}|\boldsymbol{\theta}_{ij})f\left(\boldsymbol{\eta}_{ij}\right)d\boldsymbol{\eta}_{ij}\right]f\left(\boldsymbol{\vartheta}_{j}\right)d\boldsymbol{\vartheta}_{j}, \tag{5.10}$$

where $\boldsymbol{\gamma}$ stands for the collection of freely estimated model parameters.

CHAPTER 6

Conclusion

The present research implemented a full-information multilevel item factor analysis model framework for a variety of different educational applications, with a specific emphasis on student surveys for teacher evaluation purposes. Educational data is inherently multilevel, with students nested in classrooms within schools, districts, and states. Full-information item factor analysis (IFA) models are a flexible set of models that can be used to make inferences about the latent variables (Bock et al., 1988; Mislevy, 1986; Wirth & Edwards, 2007). However, the majority of contemporary IFA models ignore the nesting of data and treat individuals and independent units. Building upon the multidimensional item-level measurement model, the multilevel item factor analysis (MLIFA) framework explicitly incorporates the nesting of individuals within level-2 units into the statistical model. Additionally, multilevel IFA models allow for the regression of latent traits on observed individual- and group-level covariates. These highly-complex models can be estimated using the Metropolis–Hastings Robbins–Monro (MH-RM) algorithm (Cai, 2010a, 2010b) implemented in flexMIRT (Cai, 2015), which allows for efficient computation of item parameters and level-1 and level-2 latent scores for a variety of item types and latent structures.

This dissertation investigated four research topics with the objectives of (a) providing an overview of multilevel IFA models applications in educational settings, (b) proposing a new approach for assessing the goodness of fit of multilevel IFA models, (c) comparing scoring approaches for producing level-2 latent score estimates in the context of relating teacher instructional practice and student achievement, and (d) examining the reliability and validity of student ratings of teacher's instructional practice. Below, I summarize the conclusions within each of these research topics.

6.1 Multilevel item factor analysis can be used to address a variety of educational topics

The broad utility of multilevel IFA models is demonstrated with a set of policy-relevant educational applications where the inference of interest is at the group-level. The three applications described in this paper are related to large-scale summative assessments and measures of teacher effectiveness, two of the most pressing and highly-debated topics in education research currently (Baker, 2016; Popham, 2013). The first application examined the estimation of school-level subscores within a Common Core State Standards (CCSS) aligned student assessment. Using a simulation study, I demonstrated that reliable school-level sub-domain scores could be estimated using the software flexMIRT under conditions similar to an existing summative mathematics assessment, provided sufficient item coverage for each sub-domain. The second application examined the classroom characteristics that are associated with positive student ratings of teacher practice. A multilevel IFA model with latent regression was estimated, and I found that classrooms with a high percentage of special education students and higher levels of student-reported effort were more likely to have positive ratings of teachers. However, ratings of teachers did not vary greatly by percentage of minority students or students' prior year test scores in English/Language Arts.

Finally, the third application was in the realm of Student Growth Percentiles (SGP), a measure of student academic growth across school years. Given that the SGPs are often aggregated to the classroom-level to produce a measure of teacher effectiveness that is used in state teacher evaluation systems, a Cluster Growth Percentile (CGP) is defined using the multilevel item factor analysis framework. The CGPs were found to be sufficiently reliable for operational use under certain circumstances.

6.2 Multilevel model fit can be addressed using a modified M_2 statistic

This dissertation sought to extend the field of limited-information goodness-of-fit assessment to the realm of multilevel item factor analysis models. This study proposed and examined the properties a limited-information statistic for multilevel IFA models. First, I demonstrated theoretically and with a simulated data example that the multilevel IFA model can be re-parameterized into a single-level item bifactor model that is fit to the group-level data. Secondly, I utilized the established relationship between the models to examine the utility of the limited-information goodness-of-fit statistic M_2 (Maydeu-Olivares & Joe, 2006) to detect misfit in re-formatted multilevel item response data. Additionally, a Reduced M_2 statistic was proposed to isolate the presence of item-level misfit. Through a series of simulation studies, I found that the existing M_2 statistic is sensitive to the examined misspecifications of the item model, but that the proposed Reduced M_2 is slightly conservative (with Type I error rates consistently below the nominal level). It is likely that the conservative nature of the Reduced M_2 is due to the degree to which the first- and second-order margins have been collapsed.

6.3 Teacher practice scores produced from the multilevel model are predictive of end-of-year test scores

This dissertation also examined the role of measurement error in attenuating the relationship between measures of teacher instructional practice and student academic growth. There were two goals of this study: (1) identify the teacher practices that are most predictive of student achievement, and (2) compare various methods for producing classroom-level latent scores that vary in the degree to which they accounted for student characteristics and uncertainty in the estimates. Specifically, this paper describes a multilevel multidimensional plausible values approach for producing classroom-level latent scores. This first stage of this approach consists of specifying and imputing sets of plausible values from a multilevel item bifactor measurement model implemented in flexMIRT (Cai, 2015). In the second stage, I fit a hierarchical linear model to predict

student achievement by the imputed teacher practice values. Using the 7Cs of teacher practice measured by Tripod survey (Ferguson, 2010), the scores from the multilevel multidimensional plausible values approach were compared with simple class means, an Expected A Posteriori (EAP) approach, and EAP with conditional means approach.

I found that overall perceptions of teacher effectiveness, as well as perceptions of a teacher's classroom management practices, were strongly related with end-of-year student academic achievement in middle school English classrooms, controlling for key student and classroom background characteristics. Additionally, I found little gain in using a multilevel plausible values approach to produce latent scores over more widely-used EAP score methods, which do not account for score uncertainty. However, gain (or lack thereof) from using plausible values is likely related to a combination of factors, including sample size, number of observed item responses used to measure the latent dimension(s), and measurement error in latent predictors. Therefore, the conclusions from this study regarding the advantages or disadvantages of the multilevel multidimensional plausible values approach do not necessarily generalize to other settings where a set of interrelated latent variables are used as predictors in a hierarchical linear model.

6.4 The Tripod survey of teacher practice is fairly reliable and related to other teacher practice measures

The Tripod Survey (Ferguson, 2010) is the most widely-used off-the-shelf student survey instrument (Bill & Melinda Gates Foundation, 2012). The Tripod student perceptions survey focuses primarily on what teachers do and how the classroom operates, which is operationalized as the Tripod 7Cs^{TM} framework of teacher effectiveness. Following the validity argument approach outlined by Kane (2006), I examined the evidence of dimensionality, bias, reliability, and concurrent validity of the Tripod survey of teacher practice.

The theoretical dimensionality of the Tripod Survey (e.g., the 7Cs of teacher practice) was not supported by the middle school student responses collected through the

Measures of Effective Teaching (MET) data. Instead, I found support for a two-factor structure, with a classroom management dimension (measured by the Control items) and a composite Support dimension consisting of the remaining six theoretical domains (e.g., Care, Clarify, Consolidate, Confer, Challenge, and Captivate). These findings are in line with the results of Ferguson's (2010) principal components analysis (PCA) using classroom item means. He found that a dimension containing Care, Clarify, Consolidate, Confer, Captivate, and Challenge accounted for half of the observed item variation, and that Control (Press) was distinguished as separate dimension. Furthermore, the importance of classroom management as a separate dimension that is predictive of student learning is supported by previous research (Evertson et al., 1983; Emmer & Stough, 2001).

The Control and Support dimensions of the Tripod survey were found to be reliable, particularly compared to widely-used measures of teacher practice. Reliability of student surveys is often summarized using a single index: Cronbach's α. This study went beyond this simple index and examined multiple criteria of reliability, including within-group agreement in the ratings of teachers within a classroom, and across-section and across-year correlation of Tripod scores. As noted by other researchers (e.g., Goldhaber & Hansen, 2013; McCaffrey, Sass, Lockwood, & Mihaly, 2009), examining the temporal stability of teacher quality implicitly assumes that job performance is a relatively stable attribute within teachers. If the dimensions of teacher practice measured by the Tripod survey tend to be unstable characteristics, then the usefulness of teacher-based accountability may be limited. In this study, the across-section and across-year correlation within a teacher were used to estimate the proportion of observed variation that is due to stable differences between teachers. The across-section correlation was over .50 for the Support and Control dimensions, implying that more than half of the observed variation is due to stable differences between teachers.

Finally, the Tripod survey was found to be related to other theoretically related measures of teacher practice. Significant correlations in the expected directions were found between the Tripod and class value-added scores, domains of the Classroom Assessment Scoring System (CLASS) observational protocols, and domains within the

Framework for Teaching observational protocol. Additionally, even after conditioning on key background variables (such as student prior achievement, race, gender) in the analysis, the Tripod Control (classroom management) domain was found to be a significant predictor of student academic achievement in both English and math.

This study provides insights into the use of the Tripod survey for high-stakes summative and formative purposes. Tripod Control and Support scores can be used to reliably distinguish among teachers at the very high and low ends of the spectrum, but most teachers scores contain overlapping confidence intervals. Therefore, it is not recommend that these scores be used to sort teachers into groups based on the middle of the distribution. Additionally, confidence intervals should be provided to aid in teacher's understandings of comparisons with other teachers or with school or district averages.

If Tripod scores are used for the purposes of prompting teachers to reflect on their practice or for directing teachers to further training, it is not recommended that separate scores be provided for Care, Clarify, Consolidate, Confer, Challenge, and Captivate. These six domains form a single dimension, and therefore the scores for each separate domain represent a less reliable estimate of the composite Support domain. The fact that the Support dimension contains multiple components that can be seen as conceptually different (e.g., Care and Challenge) does not indicate that these do not exist theoretically as separate characteristics of teachers. It merely implies that middle school students appear unable to differentiate among teachers who are Caring vs. Challenging, and generally see their teacher as either high or low on all of these six domains (Care, Clarify, Consolidate, Confer, Challenge, and Captivate).

6.5 Future directions

This study helped to demonstrate the broad range of applicability of multilevel IFA models. The results of this study suggest five areas for future research, each of which is addressed briefly below.

6.5.1 Examine parameter recovery across various conditions

I have conducted a preliminary series of simulation studies regarding the accuracy of estimation of item parameters and latent scores. However, a more systematic examination of the accuracy of MH-RM estimation across various sample sizes per group, group sample sizes, and range of ICCs would be warranted. Additionally, preliminary evidence indicates that providing accurate starting values improves the efficiency of estimation, particularly when ICCs are low and the level-2 variances are estimated in relation to fixed level-1 variances. In case of very small level-2 variances, estimation of the model may be difficult due do boundary limit on the variances (e.g., variances are constrained to be greater or equal to zero), and so parameterizations such as the multilevel "Rasch" item bifactor model used in Chapter 4 should be further investigated.

6.5.2 Expand the multilevel limited information fit statistics to include unbalanced data

The current work represents an important first step in extending the assessment of model goodness of fit to multilevel item factor analysis models. However, the current study is limited to unidimensional multilevel IFA models with balanced level-2 units. In the future, I will work on expanding the limited-information goodness of fit framework to allow for evaluating multilevel IFA models with unbalanced data.

6.5.3 Disentangle measurement error and substantive within-classroom variance

The standard approaches for modeling student survey data assumes that variance between classrooms is measuring meaningful differences in instructional quality, but that variance within classrooms represents noise (Marsh et al., 2012). Item intraclass correlations for the Tripod items in my analysis range from .18 to .32, indicating that the majority of the variance for the items is within-classrooms. If all of this within-class variance is believed to be noise, I would be forced to conclude that there is very little signal about consistent teacher practice among the noise in the ratings.

The goal of the Tripod is to measure classroom-level phenomena. However, many of the items on the Tripod survey contain wording that focus on the teacher's interaction with an individual student rather than the classroom as a whole. For example, one Tripod item is "My teacher wants me to explain my answers–why I think what I think". In this scenario, it seems implausible that all variance within the class is rater error, as the teacher is unlikely to act in the exact same manner for all students. As described by Schweig (2014), it is possible that classrooms contain micro-climates, where individual students have legitimately different experiences with instruction in a particular classroom. In reality, the within-classroom variance likely represents a combination of true differences in the perceptions of teachers, measurement error, and systematic differences in students' ratings due to differences in the use of response scales or other student characteristics.

This paper used a random intercepts model to try to explain a portion of the within-classroom disagreement. Alternative models that allow for a more systematic exploration of within classroom differences should be developed and explored.

6.5.4 Shorten the Tripod survey

The current work examined the reliability and validity of the 36-item secondary Tripod survey. Given that my analyses indicate that the full-length version of the survey provided reliable feedback, it is desirable to find ways to shorten the survey in a manner that reduces participant burden and class time loss while maintaining an acceptable level of reliability.

There are two major techniques that could be used to reduce the survey burden. The first retains all of the items, but splits the survey items across forms so that students within a class only respond to a subset of items. The second approach reduces the length of the survey so that all students are administered a shorter form that takes less time to respond to. The advantage of the first approach is that content coverage of the survey is maintained, but it assumes that all students are interchangeable and therefore all within-classroom variation is noise, which may not be a tenable assumption. The second approach is useful if the reliability of the overall score is the bigger priority over

content coverage, as a reliable overall score may be attainable using only a small set of informative items. However, this approach may not allow for reliable reporting of subscores.

Within the first approach, multiple techniques may be used to distribute items to students to reduce response burden. Items could be split into two separate forms, the two forms would be randomly assigned to students in the class. A more complex version of this is to use Balanced Incomplete Block (BIB) forms, where items are split into forms in a way that allows for an analysis of all of the items across forms. Lastly, a sample of students could be selected from each class to fill out the entire survey. In this last technique, the participant burden for those selected remains the same, but the overall class time loss is lower if the remaining students can focus on instructional activities.

For the second approach, items for a short form may be selected based on criteria related to item information or based on maximizing the relationship of the scores to some external criteria. Choosing items based on item information criteria, estimated through item factor analysis methods, would maximize the marginal reliability of the short form. However, depending on the goal of the measure, maximizing the relationship of a short form with an external criteria (e.g., a criterion measure of teacher quality) may be more important than maximizing reliability. For example, say that we are interested in creating a short form of the Tripod survey that contains the set of teacher practices items that are most strongly related to whether a student shows gains in academic achievement. Partial least squares regression, a predictive modeling technique that allows for highly collinear predictors, could be used to measure the set of items that best predict student achievement.

6.5.5 Examine the stability of the Tripod survey within a school year

According to the website of the Tripod survey, "Tripod survey results provide information that teachers can use to set specific priorities for differentiated professional development and coaching support" (Tripod Project, 2016). Since the survey is inexpensive and relatively brief to administer (particularly with shortened forms), teachers

are now being encouraged to implement the survey multiple times throughout the school year. This repeated administration allows teachers to have a chance to adapt to the feedback provided by the survey.

However, loss of classroom instruction time due to repeated teacher evaluation activities is a reasonable concern. Additionally, it is not clear that the survey is sensitive to changes in teacher practices over time in a way that would allow for professional development goal setting or monitoring. Furthermore, introducing multiple ratings may add additional noise if students rate teachers systematically differently based on time of day, day of the week, or month within the school year.

Therefore, I am interested in (a) the stability of ratings over time within a school year, and (b) sensitivity of the survey measure to changes in practice as a results of a professional development intervention. An initial study by Ron Ferguson indicated that the correlations over time (corrected for measurement error) in classroom level responses in December and March of the same school year ranged between 0.70 and 0.85 (Kane & Cantrell, 2010). To address the stability question more thoroughly, a generalizability study similar to those that have been done with classroom observations (e.g., Kane & Staiger, 2012) could be conducted. This study would examine the variance in the survey reflecting consistent differences in practice between individual teachers (as opposed to variation attributable to the student raters, or the class section being taught, or the time of year of the rating). However, this requires multiple ratings within a school year, which is not provided by the MET data, so additional data sources would need to be explored. Secondly, data could be collected alongside a professional development intervention to measure whether students using the Tripod survey respond differently before and after the intervention. These two lines of research would provide important validity evidence about the Tripod's ability to provide feedback for professional development purposes.

Bibliography

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385–456). Springer.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47–76.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment.* Author.

Asparouhov, T., & Muthén, B. O. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. In *Proceedings of the 2007 JSM meeting in Salt Lake City, Utah, Section on Statistics in Epidemiology* (pp. 2531–2535).

Baker, E. L. (2016, March). Research to Controversy in 10 Decades. *Educational Researcher*, *45*(2), 122–133.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., ... Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers.* Washington, DC: Economic Policy Institute.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: linear mixed-effects models using Eigen and S4.*

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.

Bennett, W. J. (2002). Preface: What works in teaching. In L. T. Izumi & W. F. Evers (Eds.), *Teacher Quality.* Hoover Institution Press.

Betebenner, D. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51.

Bill & Melinda Gates Foundation. (2012). *Asking Students about Teaching Student Perception Surveys and Their Implementation.* Seattle, WA.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice.* Springer Science & Business Media.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). San Francisco, CA, US: Jossey-Bass.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261–280.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. Witrock (Ed.), *The third handbook of research on teaching* (pp. 328–375). New York: Macmillan.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (Vol. 154, pp. 136–162). Newbury Park, CA: Sage.

Bush, G. W. (2001). *No Child Left Behind* (No. 10-110). Washington, DC: US Department of Education.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Cai, L. (2015). *flexMIRT version 3: Flexible multilevel multidimensional item analysis and test scoring.* Seattle, WA: Vector Psychometric Group.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 245–276.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2P tables. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 173–194.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological methods*, *16*(3), 221.

Camburn, E. M. (2012, November). *Review of "Asking Students about Teaching".* Boulder, CO: National Education Policy Center.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, *44*(5), 495–518.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*(2), 189–225.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (Working paper No. 17699). Cambridge, MA: National Bureau of Economic Research.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching.* ASCD.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8–15.

De Jong, M. G., Steenkamp, J.-B. E., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of marketing research*, *45*(1), 104–115.

Doherty, K., & Jacobs, S. (2013). *State of the States 2013 Connect the Dots: Using evaluations of teacher effectiveness to inform policy and practice.* National Council on Teacher Quality.

Doherty, K., & Jacobs, S. (2015). *State of the States 2015: Evaluating Teaching, Leading and Learning.* National Council on Teacher Quality.

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational psychologist*, *36*(2), 103–112.

Evertson, C. M., Emmer, E. T., Sanford, J. P., & Clements, B. S. (1983). Improving classroom management: An experiment in elementary school classrooms. *The Elementary School Journal*.

Falk, C. F., & Cai, L. (2015). A Flexible Full-Information Approach to the Modeling of Response Styles. *Psychological Methods*.

Ferguson, R. F. (2010). *Student perceptions of teaching effectiveness* (Discussion brief from the National Center for Teacher Effectiveness and the Achievement Gap Initiative). Boston, MA: Harvard University.

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, *94*(3), 24–28.

Ferguson, R. F., & Danielson, C. (2014). How Framework for Teaching and Tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project.* San Francisco, CA: Jossey-Bass.

Fox, J.-P. (2010). Multilevel Item Response Theory Models. In *Bayesian Item Response Modeling* (pp. 141–191). Springer New York.

Fox, J.-P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*(2), 271–288.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis.

Psychometrika, *57*(3), 423–436.

Goldhaber, D., & Hansen, M. (2013). Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance. *Economica*, *80*(319), 589–612.

Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way error components analysis of educational productivity. *Education Economics*, *7*(3), 199–208.

Goldstein, H., & Steele, F. (2005). Multilevel factor analysis models for continuous and discrete data. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary Psychometrics. A Festschrift to Roderick P. McDonald.* Mahwah, NJ: Lawrence Erlbaum.

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores* (14th William H. Angoff Memorial Lecture). Princeton, NJ: ETS.

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014, May). Evidence for General and Domain-Specific Elements of Teacher–Child Interactions: Associations With Preschool Children's Development. *Child Development*, *85*(3), 1257–1274.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. *Handbook of the Economics of Education*, *2*, 1051–1078.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal*, *105*(1), 11–30.

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Joe, J. N., Tocci, C. M., Holtzman, S. L., & Williams, J. C. (2013). Foundations of observation: Considerations for developing a classroom observation system that helps districts achieve consistent and accurate scores. *MET Project Policy and Practice Brief. Seattle, WA: Bill & Melinda Gates Foundation*.

Kamata, A. (2001). Item Analysis by the Hierarchical Generalized Linear Model. *Journal of Educational Measurement*, *38*(1), 79–93.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger Publishers.

Kane, T. J., & Cantrell, S. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project.* Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment (MET Project Research Paper).* Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation.* National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. (MET Project Research Paper).* Seattle, WA: Bill & Melinda Gates Foundation.

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis*, *21*(4), 345–363.

Kuhfeld, M., Cai, L., & Monroe, S. L. (in preparation). Full-information multilevel item factor analysis with applications.

Lee, V. E., Smith, J. B., Perry, T. E., & Smylie, M. A. (1999). *Social Support, Academic Press, and Student Achievement: A View from the Middle Grades in Chicago. Improving Chicago's Schools. A Report of the Chicago Annenberg Research Project.* Chicago, IL: Consortium on Chicago School Research.

Liaw, S.-H., & Goh, K.-L. (2003). Evidence and control of biases in student evaluations of teaching. *International Journal of Educational Management*, *17*(1), 37–43.

Lockwood, J. R., & Castellano, K. E. (2015). Alternative statistical frameworks for student growth percentile estimation. *Statistics and Public Policy*, *2*(1), 1–9.

Lu, I. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling*, *12*(2), 263–277.

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.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International journal of educational research*, *11*(3), 253–388.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom Climate and Contextual Effects: Conceptual and Methodological Issues in the Evaluation of Group-Level Effects. *Educational Psychologist*, *47*(2), 106–124.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological methods*, *11*(4), 344.

McCaffrey, D. F. (2012, October). *Do value-added methods level the playing field for teachers?* Carnegie Knowledge Network.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education*, *4*(4), 572–606.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational and Behavioral Statistics*, *11*(1), 3–31.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161.

Monroe, S., & Cai, L. (2015). Examining the Reliability of Student Growth Percentiles Using Multidimensional IRT. *Educational Measurement: Issues and Practice*, *34*(4), 21–30.

Muthén, B., & Asparouhov, T. (2013). Item response modeling in Mplus: A multidimensional, multi-level, and multi-timepoint example. In W. Van Der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory: Models, statistical tools, and applications.* Chapman & Hall.

Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. Hox & J. K. Roberts (Eds.), *Handbook of Advanced Multilevel Analysis* (pp. 15–40). New York: Taylor and Francis.

National Center for Research on Evaluation, Standards, and Student Testing. (2015, June). *Simulation-Based Evaluation of the Smarter Balanced Summative Assessments.* Los Angeles, CA: University of California, Los Angeles.

National Governors Association. (2014). *Trends in state implementation of the common core state standards: Making the shift to better tests.* Washington, DC: National Governors Association.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational evaluation and policy analysis*, *26*(3), 237–257.

Phillips, M., Yamashiro, K., Schaaf, K., & Schweig, J. (2011). *The Los Angeles Unified School District pilot of Classroom and School Environment surveys: A technical report exploring reliability and validity in nine School Improvement Grant (SIG) schools.* Los Angeles, CA: Los Angeles Education Research Institute.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational researcher*, *38*(2), 109–119.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system.* Baltimore, MD: Brookes.

Popham, W. J. (2013). *Evaluating America's Teachers: Mission Possible?* Corwin.

R Core Team. (2012). *R: A language and environment for statistical computing* (Computer software manual). Vienna, Austria: R foundation for Statistical Computing.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190.

Raudenbush, S., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Raudenbush, S., & Jean, M. (2014). To What Extent Do Student Perceptions of Classroom Quality Predict Teacher Value Added? In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project* (pp. 170–202). San Francisco, CA: Jossey-Bass.

Rijmen, F., Jeon, M., von Davier, M., & Rabe-Hesketh, S. (2013). A general psycho-

metric approach for educational survey assessments: Flexible statistical models and efficient estimation methods. In D. Rutkowski & M. von Davier (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 583–606).

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *The American Economic Review*, *94*(2), 247–252.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). *Can you recognize an effective teacher when you recruit one?* (Working Paper 14485). Cambridge, MA: National Bureau of Economic Research.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the Study of Instructional Improvement. *Educational Researcher*, *38*(2), 120–131.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement*, *17*(4).

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

Schweig, J. D. (2014). *Multilevel factor analysis by model segregation: Comparing the performance of maximum likelihood and robust test statistics* (Unpublished doctoral dissertation). University of California, Los Angeles, Los Angeles, CA.

Shang, Y., VanIwaarden, A., & Betebenner, D. W. (2015). Covariate measurement error correction for student growth percentiles using the SIMEX method. *Educational Measurement: Issues and Practice*, *34*(1), 4–14.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575.

Skrondal, A., & Rabe-Hesketh, S. (2005). Structural equation modeling: categorical variables. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 4, pp. 1905–1910). John Wiley & Sons.

Smarter Balanced Assessment Consortium. (2013). *Initial Achievement Level escriptors and College Content-Readiness Policy.* https://www.smarterbalanced.org/wp-content/uploads/2015/.

Smarter Balanced Assessment Consortium. (2016). *What is Smarter Balanced?* http://www.smarterbalanced.org/about/.

Stucky, B. D., Thissen, D. M., & Edelen, M. O. (2012, November). Using Logistic Approximations of Marginal Trace Lines to Develop Short Assessments. *Applied Psychological Measurement*, 1-17.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: a search for truth or a witch hunt in student ratings of instruction? *New directions for institutional research*, *2001*(109), 45–56.

Toch, T., & Rothman, R. (2008). Rush to Judgment: Teacher Evaluation in Public Education. Education Sector Reports. *Education Sector*.

Tripod Project. (2016). *Become a Tripod School.* http://tripoded.com/school-leaders/.

Tucker, L. R. (1971, December). Relations of factor score estimates to their use. *Psychometrika*, *36*(4), 427–436.

U.S. Department of Education. (2009). *Race to the Top Program Executive Summary.* Washington, DC: U.S. Department of Education.

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI monograph series*, *2*, 9–36.

von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174).

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* Cambridge University Press.

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

White, M., & Rowan, B. (2013, October). *Measures of Effective Teaching: Grantee*

Files, 2009-2011 (Grantee User Guide). Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological methods*, *12*(1), 58.

Yang, J. S., & Seltzer, M. (2016). Handling measurement error in predictors with a multilevel latent variable plausible values approach. In S. N. Beretvas, J. Harring, & L. Stapleton (Eds.), *Advances in Multilevel Molding for Educational Research: Addressing Practical Issues Found in Real-World Applications.* Charlotte, NC: Information Age Publishing, Inc.