Fall 1-6-2017

The Variation of a Teacher's Classroom Observation Ratings across Multiple Classrooms

Xiaoxuan Lei

ACCEPTANCE

This dissertation, THE VARIATION OF A TEACHER'S CLASSROOM OBSERVATION RATINGS ACROSS MULTIPLE CLASSROOMS, by XIAOXUAN LEI, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree, Doctor of Philosophy, in the College of Education and Human Development, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chairperson, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty.

____
Hongli Li, Ph.D.
Committee Chair

____
Audrey Leroux, Ph.D.
Committee Member

____
Kevin Fortner, Ph.D.
Committee Member

____
Lee Branum-Martin, Ph.D.
Committee Member

____
Date

____
William Curlette, Ph.D.
Chairperson, Department of Educational Policy
Studies

____
Paul Alberto, Ph.D.
Dean, College of Education and Human
Development

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education and Human Development's Director of Graduate Studies, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

____

XIAOXUAN LEI

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Xiaoxuan Lei
Educational Policy Studies
College of Education and Human Development
Georgia State University

The director of this dissertation is:

Dr. Hongli Li
Department of Educational Policy Studies
College of Education and Human Development
Georgia State University
Atlanta, GA 30303

CURRICULUM VITAE

Xiaoxuan Lei

EDUCATION:

Ph.D.	2016	Georgia State University/ Department of Educational Policy Studies Research, Measurement, and Statistics
Master's Degree	2012	Florida State University/College of Education Higher Education Administration

PROFESSIONAL EXPERIENCE:

2015-2016	Research Assistant Center for the Study of Adult Literacy
2014-2015	Research Assistant Language & Literacy Initiative Seed Grant
2014-2015	Research Assistant Quality English and Science Teaching Project
2013-2015	Research Assistant Developmental Psychology Program
2012-2013	Research Assistant Network for Enhancing Teacher Quality Project

PRESENTATIONS AND PUBLICATIONS:

Lei, X., Branum-Martin, L., Taylor, P., Carlson, C. D., & Francis, D. J. (2015, July). *Quantity versus quality of reading instruction.* Paper presented at the 22nd annual meeting of the Society for the Scientific Study of Reading, Waimea, HI.

Li, H, Fortner, C. K., Qi, Q., Lei, X. (2015, April). *An Examination of Teachers' Assessment Practices in the US: Evidence from the TIMSS*. Paper was presented at the American Educational

Research Association (AERA), Chicago, IL.

Branum-Martin, L., Mehta, P. D., Taylor, W. P., Carlson, C. D., Alic, X., Hunter, C. V., & Francis, D. J. (2015, March). *How do we match instructional effectiveness with learning curves?* Paper was presented at the Society for Research on Educational Effectiveness, Washington, D.C.

Lei, X., & Li, H. (2014, April). *The effect of what 15-year-old US students read for school on their performance on reading continuous and non-continuous texts in PISA 2009*. Paper presented at the American Educational Research Association (AERA), Philadelphia, PA.

Lei, X., & Chen, J. (2014, April). *The impact of bilingual education funding and other factors on academic achievement*. Paper presented at the American Educational Research Association (AERA), Philadelphia, PA.

Li, H., Fortner, C. K., & Alic, X. (2014). Relationships between the Use of Test Results and U.S. Students' Academic Performance. *School Effectiveness and School Improvement*, *26*(2), 258-278. doi: 10.1080/09243453.2014.898662

PROFESSIONAL SOCIETIES AND ORGANIZATIONS

2012-2016 American Educational Research Association (AERA)

THE VARIATION OF A TEACHER'S CLASSROOM OBSERVATION RATINGS ACROSS MULTIPLE CLASSROOMS

by

XIAOXUAN LEI

Under the Direction of Dr. Hongli Li

ABSTRACT

Classroom observations have been increasingly used for teacher evaluations, and thus it is important to examine the measurement quality and the use of observation ratings. When a teacher is observed in multiple classrooms, his or her observation ratings may vary across classrooms. In that case, using ratings from one classroom per teacher may not be adequate to represent a teacher's quality of instruction. However, the fact that classrooms are nested within teachers is usually not considered while classroom observation data is analyzed. Drawing on the Measures of Effective Teaching dataset, this dissertation examined the variation of a teacher's classroom observation ratings across his or her multiple classrooms. In order to account for the teacher-level, school-level, and rater-level variation, a cross-classified random effects model was used for the analysis. Two research questions were addressed: (1) What is the variation of a teacher's classroom observation ratings across multiple classrooms? (2) To what extent is the classroom-level variation within teachers explained by observable classroom characteristics?

The results suggested that the math classrooms shared 4.9% to 14.7% of the variance in the classroom observation ratings and English Language and Arts classrooms shared 6.7% to 15.5% of the variance in the ratings. The results also showed that the classroom characteristics (i.e., class size, percent of minority students, percent of male students, percent of English language learners, percent of students eligible for free or reduced lunch, and percent of students with disabilities) had limited contributions to explaining the classroom-level variation in the ratings. The results of this dissertation indicate that teachers' multiple classrooms should be taken into consideration when classroom observation ratings are used to evaluate teachers in high-stakes settings. In addition, other classroom-level factors that could contribute to explaining the classroom-level variation in classroom observation ratings should be investigated in future research.

INDEX WORDS: Classroom observation ratings, Classroom-level variation, Cross-classified random effects modeling

THE VARIATION OF A TEACHER'S CLASSROOM OBSERVATION RATINGS ACROSS MULTIPLE CLASSROOMS

by

XIAOXUAN LEI

A Dissertation

Presented in Partial Fulfillment of Requirements for the

Degree of

Doctor of Philosophy

in

RESEARCH, MEASUREMENT, AND STATISTICS

in

EDUCATIONAL POLICY STUDIES

in the

College of Education and Human Development

Georgia State University

Atlanta, GA
2016

Table of Contents

LIST OF TABLES

LIST OF FIGURES

1 THE PROBLEM

Background

Recent research indicates that the teacher is a very important factor affecting student learning outcomes (Chetty, Friedman, & Rockoff, 2011; Darling-Hammond, 2000; Rockoff, 2004; Sanders, Wright, & Horn, 1997; Staiger & Rockoff, 2010). Therefore, the past decades have seen federal legislation put states and districts under pressure to improve and evaluate teacher quality. In 2002, the No Child Left Behind (NCLB) Act was enacted to help the nation's students increase the academic achievement by improving school and teacher quality. One of the goals of the NCLB for states and districts was to recruit and prepare "highly qualified" teachers to support students' academic achievement (U.S. Department of Education, 2015a). Supported by the NCLB and the Race to the Top (RTTT) funding, a competitive grant for rewarding reforms in state and district K-12 education, over two-thirds of states have upgraded their teacher evaluation systems by incorporating student achievement data as a measure of teacher effectiveness alongside other measures, such as classroom observations and student surveys since 2009 (Hull, 2013). In December 2015, a new educational law, The Every Student Succeeds Act (ESSA), was signed by President Obama, which emphasized providing assistance to local education agencies to support the design and implementation of teacher evaluation with multiple measures of educator performance (U.S. Department of Education, 2015b). Driven by the trend of educational policy in teacher evaluation systems in the past decades, states have made efforts to build their teacher evaluation systems using multiple methods to measure teacher effectiveness (McGuinn, 2012; Partee, 2012; Steele, Hamilton, & Stecher, 2010).

Among teacher effectiveness measures, value-added models (VAMs) are popular statistical models available for measuring teacher effectiveness using student achievement data

(Hull, 2013). VAMs use students' prior achievement on standardized tests to predict their achievement in the next year and produce effect estimates on growth attributable to teachers and schools rather than to other sources (Geo, Bell, & Little, 2008; Lockwood, Louis, & McCaffrey, 2002). An assumption underlying the use of VAMs is that teachers whose students have higher value-added scores are providing better instruction than teachers whose students have lower scores (Marzano & Toth, 2013). Along with the use of VAMs, classroom observation is another important component of states' teacher evaluation systems (Hull, 2013; National Council of Teacher Quality, 2015; Whitehurst, Chingos, & Lindquist, 2014). Teachers are usually observed multiple times a year by trained evaluators using a rubric. Some researchers (Mihaly & McCaffrey, 2014; National Council on Teacher Quality, 2015; Polikoff, 2015) believe that classroom observation is a relatively accurate measure in judging classroom instructional practices.

Classroom observation is a method of measuring classroom behaviors from direct observations or recorded observations. The data collected from this procedure is usually based on coding the frequency or quality of specific behaviors between students and teachers occurred in the classroom during a given time interval (Board, 2011; Waxman & Huang, 1999). Classroom observation ratings have been used as standard-based evaluations of practice to measure teachers' classroom performance (Darling-Hammond, 2012). Some states (e.g., Arizona, Utah) applied observation ratings to as much as 40 to 75 percent of the total scores in their teacher evaluation systems for high-stakes decision-making in tenure, promotion, and compensation (Partee, 2012; Whitehust et al., 2014). As an illustration, the Hillsborough County Public School District in Florida implemented the Empowering Effective Teachers program in 2010-2011 academic year with 60 percent of each teacher's performance evaluation based on classroom observations (Steele et al., 2010). Additionally, some school districts used teacher evaluation scores consisting of a weighted combination of classroom observation ratings and other teacher effectiveness measures (Hansen, Lemke, & Sorensen, 2013; Kane & Staiger, 2012; Leo & Lachlan-Haché, 2012; Mihaly, McCaffrey, Staiger, & Lockwood, 2013). Moreover, classroom observation ratings also have the potential of providing formative feedback to help teachers improve their teaching practices (Darling-Hammond, 2010; Hill et al., 2012; Whitehurst et al., 2014).

The U.S. Department of Education has envisioned equitable and transparent teacher evaluation systems with multiple measurements of teacher effectiveness in states and districts that inform compensation, tenure, and dismissal (U.S. Department of Education, 2015c). In addition, ESSA encouraged states and districts to develop plans to improve the quality of teacher evaluation such as developing classroom observation rubrics and methods for ensuring the reliability and validity of evaluation results (Partee, 2012; U.S. Department of Education, 2015b). However, one important consideration for states and districts is whether the classroom observation ratings can adequately represent a stable characteristic of teaching quality for a specific teacher. Classroom observations, as a sampling of classroom behavior over time, may be subject to several sources of systematic variation, which could affect the ratings and bias the evaluation results attributed to teachers (Kennedy, 2010). Many observational systems evaluate a teacher's classroom performance using multiple raters per teacher, a sample of the teachers' multiple lessons, and a sample of the teachers' instruction from multiple times (Kelcey, McGinn, & Hill, 2013). A teacher's observation ratings may vary across these occasions. If classroom observation ratings are used as standard-based evaluations regarding teachers, the sources of systematic variation could be construct-irrelevant factors (Kelcey et al., 2013; Kennedy, 2010).

For example, if a teacher’s observation ratings fluctuate greatly from time to time, using one-time observation to evaluation the teacher may not be accurate. According to Morgan, Hodge, Trepinski, and Anderson (2014), “the desirability of stability is largely a function of the purpose for which the data are to be used” (p. 4). For example, for employment or promotion decisions, a stability of teacher quality measurement is important (i.e., consistently poor or high); for compensation decisions, desirability of stability may link to one particular occasion (e.g., matching teachers with their students, grade levels, or subjects at a particular time) (Morgan et al., 2014). Therefore, it is important to examine the variation of classroom observation ratings across various construct-irrelevant factors (e.g., times, raters) for the interpretation and use of teacher effectiveness measures.

Prior research demonstrated that a teacher’s observation ratings showed variation across different occasions. For example, Hill et al. (2012) found that a teacher’s observation ratings of the Mathematical Quality of Instruction (MQI) were not constant across the lessons he or she taught. Bell et al. (2012) also found that a teacher’s observation ratings of the Classroom Assessment Scoring System for the secondary classrooms (CLASS-S) were not constant across the lessons he or she taught. Furthermore, Smolkowski, and Gunn (2012) showed that a teacher’s observation ratings of the Classroom Observations of Student-Teacher Interactions (COSTI) were not constant across different times that he or she was observed. Polikoff (2015) also found that a teacher’s observation ratings were not stable across years. However, these studies used data from one classroom per teacher without examining the variation of a teacher’s classroom observation ratings across his or her multiple classrooms. When a teacher teaches multiple classrooms, it should not be assumed that his or her classroom observation ratings are stable across their classrooms. According to Bell et al. (2012), “teaching occurs in a context and is inextricably tied to aspects of that context" (p. 85). However, in almost all of the research evidence, the teacher was not disentangled from the classroom as a teaching context. That is, teachers might be assigned classroom observation ratings no matter what type of classrooms or groups of students they taught when they were observed. In this case, it may be imprudent to make high-stakes decisions for a teacher as "excellent" for one group of students, and "medium" for another group of students.

In a research report of the Measures of Effective Teaching (MET) project, Kane and Staiger (2012) decomposed the total variance in classroom observation ratings into various components including teachers, classrooms, lessons, raters, and their interactions. The results showed that classrooms in general explained 0 to 11 percent of the variance in the classroom observation ratings depending on the observation instrument. However, Kane and Staiger's (2012) analysis was rather general, where aggregated ratings across video segments were used, the school-level variation was ignored, coarse observation outcome variables were used, and subject differences were not considered (see more details in the Purpose of the Study section). In particular, they did not attempt to explain the classroom-level variation in observation ratings. Using ratings of the Classroom Assessment Scoring System (CLASS; Pianta, La Paro & Hamre, 2008) collected by the MET project, this dissertation provided a further analysis and explored the variation of a teacher's classroom observation ratings across multiple classrooms as a function of classroom characteristics.

Problem Statement

Classroom observation instruments usually focus on measuring specific interactions between students and teachers in the classroom (Board, 2011; Waxman & Huang, 1999). The teaching quality in the classroom that is measured and calculated could be both the teacher's performance and the classroom's effects in response to the complexity of classrooms (Berliner, 2014). However, when classroom observation ratings are used as standard-based evaluations regarding teachers, teachers may need to be detangled from classrooms as teaching contexts. If a teacher's observation ratings are constant across all the classrooms he or she teaches, ratings from any of the classrooms can be representative of his or her teaching performance for personnel decisions (Bell et al., 2012; Kane, 2006). However, if a teacher's observation ratings are not constant across his or her classrooms, measures of teacher effectiveness from a single classroom may not be appropriate for making operational decisions regarding teachers (Kennedy, 2010). Therefore, this dissertation examined the variation of a teacher's observation ratings across multiple classrooms for the interpretation and use of classroom observation ratings.

Moreover, instruments may not be pure measures of teacher quality and the validity of instruments may be sensitive to contextual features (Bell et al., 2012). The classroom-level variation of a teacher's observation ratings may be reflected by the features of the classroom context, such as the demographic characteristics of students in the classroom and the class size (Bell et al., 2012; Darling-Hammond, 2012; Whitehurst et al. 2014). Additionally, Whitehurst et al. (2014) suggested that a statistical adjustment of classroom observation ratings for student demographics in the classroom is successful in producing a new pattern of teachers' ratings. Thus, another important question to consider is to what extent the classroom characteristics contribute to the classroom-level variation in a teacher's classroom observation ratings.

Purpose of the Study and Research Questions

The purpose of this dissertation was to examine the variation of a teacher's classroom observation ratings across multiple classrooms as a function of classroom characteristics using data collected by the Measures of Effective Teaching (MET) project. The MET researchers collected a variety of measures regarding teaching quality in classrooms over a two-year period (Academic Year 2009-2010 and 2010-2011) in the United States. More than 2,500 teachers in grades four through nine participated in the study (White & Rowan, 2013).

Typical multilevel modeling can be applied to a purely hierarchical data structure where the first level units are clustered by only one type of higher-level unit, for example, students are clustered by schools (Raudenbush & Bryk, 2002). However, in the MET dataset, ratings at the first level are not clustered by one single type of higher-level unit. Instead, ratings at the first level are clustered by more than two types of higher-level units. Ratings are cross-classified by raters and classrooms within teachers within schools, while raters and classrooms are not clustered by each other (see more details in the Review of the Literature and the Methodology sections). Modeling this type of cross-classified data structure using typical multilevel modeling may generate biased estimates (Luo & Kwok, 2010; Meyers & Beretvas, 2006; Wallace, 2015). Therefore, a cross-classified random effects model (CCREM; Goldstein, 2003; Raudenbush & Bryk, 2002) was used to handle this type of cross-classified data structure in this dissertation.

Primarily, this dissertation examined the classroom-level variation within teachers using the classroom observation ratings from teachers who taught two classrooms in the MET project. Second, how classroom observation ratings vary across their classrooms due to the classroom characteristics, such as class size and classroom composites, was examined.

Two questions were addressed in this dissertation:

1. What is the variation of a teacher’s classroom observation ratings across multiple classrooms?
2. To what extent is this classroom-level variation within teachers explained by observable classroom characteristics?

The analysis conducted in this dissertation is different from the one conducted by Kane and Staiger (2012) in a number of ways. First, Kane and Staiger (2012) calculated the variance at each level without explaining the potential causes of the variance. This dissertation examined how the classroom characteristics, including class size and classroom composites, explained the classroom-level variation in observation ratings.

Second, Kane and Staiger (2012) used domain ratings averaged across dimension ratings as the outcome measures. The CLASS instrument has three broad domains of measurement (Emotional Support, Classroom Organization, and Instructional Support) with several dimensions belonging to each domain. For example, the dimensions of Behavior management, Productivity, and Instructional learning formats that are subscales of the CLASS instrument belong to the domain of Classroom Organization. The dimension ratings can be aggregated into the domain ratings as the outcome measures. However, using domain ratings averaged across dimension ratings may lose important information. Primarily, the dimensions describe the features of teachers' performance in the classroom in more specific ways than the broader domains (The National Center on Quality Teaching and Learning, 2012). Ratings on each dimension can provide teachers and policy-makers with more actionable information for improving professional development or understanding program progress (The National Center on Quality Teaching and Learning, 2012). Additionally, previous statistical analyses showed that the theoretical three-factor model of the CLASS only moderately fit with the original twelve dimensions (Hamre, Pianta, Mashburn, & Downer, 2007; Pakarinen et al., 2010; Sandilos, DiPerna, & The Family Life Project Key Investigators, 2014; Yuan, McCaffrey, & Savitsky, 2013). These results challenged the validity of using the three domain ratings of the CLASS.

Therefore, this dissertation used ratings on each dimension of the CLASS as the outcome measures instead of the averaged ratings on each domain.

Third, in Kane and Staiger (2012), the outcome ratings of each domain were the averaged video scores across segments. In the MET project, each video taken from the classrooms was divided into two 15-minute segments, where raters scored each of these segments based on the CLASS rubrics (White & Rowan, 2013). Kane and Staiger (2012) aggregated the values of segment-level units into fewer values of video-level units. As a result, important information could have been lost due to this aggregation procedure (Hox, Moerbeek, & van de Schoot, 2010). Therefore, this dissertation used segment ratings of each dimension as the outcome variables in the analysis.

Furthermore, Kane and Staiger (2012) did not control for the school-level variation in their analysis. Prior research showed that school-level characteristics had associations with teaching quality measured by classroom observations (Abbott & Fouts, 2003; Cadima, Peixoto, & Leal, 2014). Teachers in the MET project were from many schools, and teachers working in the same school shared the common environment and policy. Thus, classroom-level variation within teachers from multiple schools may be different due to different school contexts. Moreover, ignoring a level of nesting in a multilevel analysis can impact the estimates of variance components and fixed effects, and the standard error coefficients of the lower level variables will generally be smaller resulting in inflated Type I error rates (Hox et al., 2010; Raudenbush & Bryk, 2002). Therefore, in order to account for the higher-level contexts of teachers and classrooms, this dissertation controlled for the school-level variation in the statistical analyses.

Finally, differently from Kane and Staiger (2012), the two subjects of English Language and Arts (ELA) and mathematics were analyzed separately in this dissertation. Hill et al. (2012) suggested researchers should examine whether a general instrument intended for use across academic subjects performs equally well on all subjects. ELA and math are two different subjects that may lead to different interactions between teachers and students. Furthermore, Polikoff (2015) analyzed the year-to-year stability of classroom observation ratings separately for ELA and math. Results showed that the year-to-year stability of the CLASS dimension ratings in ELA was generally lower than in math across dimensions. Thus, it is possible that the classroom-level variation in the CLASS dimension ratings may be different between ELA and math.

In conclusion, the purpose of this dissertation was to examine the variation of a teacher's classroom observation ratings across multiple classrooms as a function of classroom characteristics. The segment-level ratings on each dimension of the CLASS instrument were used as the outcome variables, and the two subjects (i.e., math and ELA) were analyzed separately using a CCREM.

Significance of the Study

This dissertation has implications for the interpretation and use of classroom observation ratings in teacher evaluations. First, if a teacher's classroom observation ratings fluctuate from classroom to classroom, he or she could be wrongly classified in teacher evaluations based on the observation ratings from only one of the classrooms. When classroom observation ratings are used for high-stakes decisions regarding teachers, researchers and evaluators may need to take the classroom context into consideration. Instead of comparing different teachers' classroom observation ratings, this dissertation compared the observation ratings from multiple classrooms of the same teacher. The classroom-level variation in observation ratings within teachers indicates how much the classrooms contribute to the variation of observation ratings instead of teachers. The second question of this dissertation (i.e., to what extent the variation of a teacher's classroom observation ratings across multiple classrooms is explained by classroom characteristics) could identify the potential classroom-level factors that can be used for the observation rating adjustment in teacher evaluation systems. This examination has implications regarding how classroom observation ratings can be used for teacher evaluations.

Summary

Due to the importance of building reliable teacher evaluation systems, there is a growing need to examine the measurement quality of classroom observation ratings as an important measure of teacher effectiveness. Classroom observation ratings can be influenced by several sources of systematic factors (e.g., lesson, rater) that can affect the validity of the ratings and bias the results attributed to teachers. The classroom-level variation in teachers' classroom observation ratings was usually not included in prior research studies. The purpose of this dissertation was to examine the variation of classroom observation ratings across multiple classrooms as a function of classroom characteristics. It is important to examine this problem for the interpretation and use of classroom observations in teacher evaluations. A review of the literature is presented in Chapter 2, including the conceptual framework of investigating the variation in classroom observation ratings across a teacher's multiple classrooms. Additionally, an introduction of the CCREM is presented in Chapter 2.

2 REVIEW OF THE LITERATURE

One important consideration for policy-makers and researchers to support states and districts in implementing reliable teacher evaluations is whether the classroom observation ratings can adequately represent a stable characteristic of teaching quality for a specific teacher. Chapter 2 begins with a conceptual framework on the hypothesis that a teacher's classroom observation ratings may vary across his or her multiple classrooms as a function of a set of classroom contextual factors. Further, this chapter follows by an introduction of the cross-classified random effects model (CCREM) that was used in the analyses of this dissertation.

Classroom-Level Variation in Teachers' Classroom Observation Ratings

Conceptual framework

The growing use of classroom observation instruments raises the issue of the degree to which classroom observation ratings represent the underlying construct the items seek to measure (Bell et al., 2012; Hill et al., 2012). Thus, it is important to understand whether the sample of teaching behaviors observed is representative of all the instances of teaching over the conditions of observation (e.g., multiple raters, multiple times) (Bell et al., 2012). As stated by Bell et al. (2012), "it is important to note that the integration between teaching quality and the contextual features of classrooms means that measures of teaching quality necessarily capture aspects of context" (p. 65). In other words, the classroom environment may influence the quality of interactions between students and teachers measured by classroom observations (Bell et al., 2012; Darling-Hammond, 2012). Therefore, it is possible that teaching quality measured by classroom observations may vary across teachers' different classrooms due to the characteristics of the classroom as contextual factors, such as students assigned to the class and class size.

Bell et al. (2012) proposed a teaching quality framework, which is illustrated in Figure 1. As shown in Figure 1, teaching quality consists of six constructs. Classroom observations measure two of the constructs, which are teacher practices and student practices. In addition, teaching quality could be influenced by contextual factors, which refer to the curriculum being used, the building leadership that supports teaching, students and colleagues, resources, and other related school and classroom characteristics. This framework displayed in Figure 1 adds weight to the argument that observation ratings may be influenced by teacher performance and other aspects of the observational environment, including students assigned to the classroom (Hill et al., 2012). Therefore, it is possible that if a teacher has multiple classrooms, his or her classroom observation ratings may vary across his or her different classrooms.

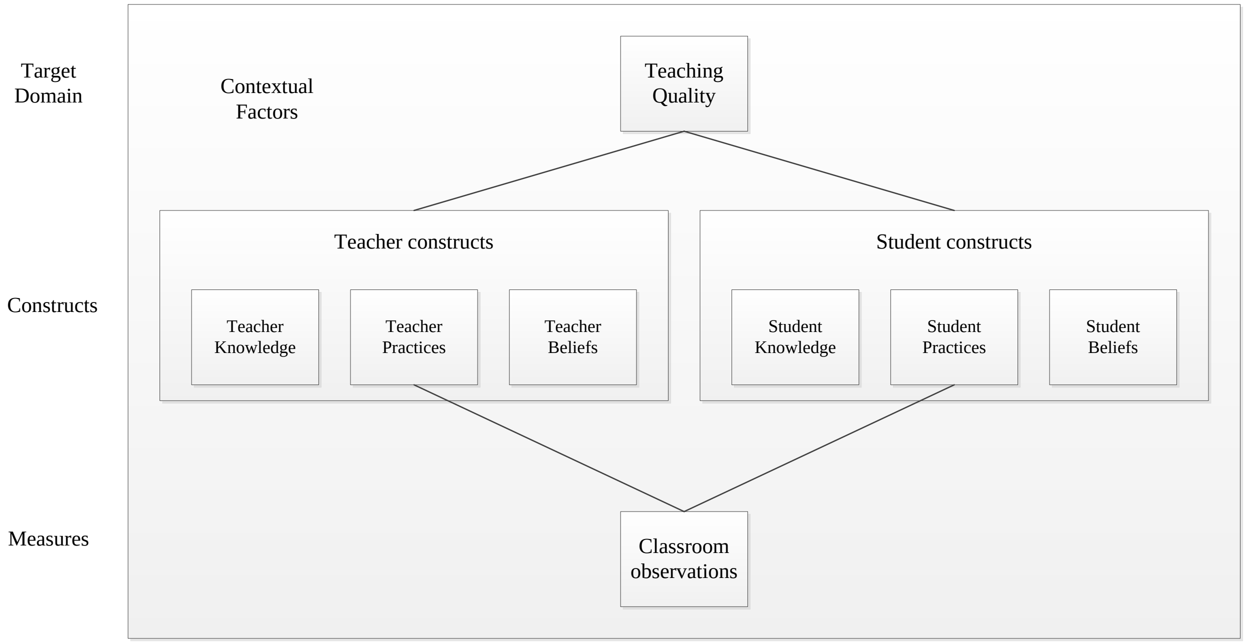

Figure 1. Conceptualizing teaching quality, contextual factors, and classroom observations. Adapted from "An Argument Approach to Observation Protocol Validity," by C. A. Bell, D. H. Gitomer, D. F. McCaffrey, B. K. Hamre, R. C. Pianta, and Y. Qi, 2012, *Educational Assessment, 17*(2-3), p. 64.

Classroom characteristics

In order to investigate how classroom observation ratings vary as a function of classroom characteristics, it is important to add related variables to examine how these predictors explain the variance at the classroom level. Whitehurst et al. (2014) found that adjusting the observation scores by controlling for the student achievement level in classrooms could move some teachers out of their original ranking positions in teacher evaluations. However, in the sample of the Measures of Effective Teaching (MET) project, different state tests were administered to students depending on six districts, two subject areas (i.e., math and ELA), and six grade levels (i.e., 4th, 5th, 6th, 7th, 8th, and 9th). In this case, 72 (i.e., 2 x 6 x 6) different tests were involved in the sample. White and Rowan (2014) cautioned researchers that student state test scores in the MET project were converted to rank-based *z*-scores within district, subject, and grade. That is, each student's *z*-score was relative to other students' *z*-scores in that particular district, subject, and grade. Therefore, the student achievement level was not used as a predictor for the datasets involving all the six districts and six grades in this dissertation. However, there are circumstances that adjusting for student achievement level is not possible (Whitehurst et al., 2014). For example, student achievement scores are not available for non-tested grades and subjects. As suggested by Whitehurst et al. (2014), this problem can be solved by controlling for student composites in the classrooms. Therefore, it is also important to explore to what extent the student composites in classrooms can explain the classroom-level variation in classroom observation ratings.

In this dissertation, it was expected that after the classroom characteristics were added, the variation of a teacher's observation ratings across classrooms might appear less. Classroom characteristics could be measured by contextual characteristics such as class size (Marsh et al., 2012) and compositional characteristics such as student composition (Dreeben & Barr, 1988; Hattie, 2002). Previous studies showed that classrooms with fewer students led to better learning and classroom processes (Blatchford, Bassett, & Brown, 2011; Bruhwiler & Blatchford, 2011; Curby et al., 2011; La Paro et al., 2009). Thus, class size was used as one of the classroom-level predictors in the analysis. Furthermore, Polikoff (2015) used classroom demographic characteristics as the predictors to explain the variation of classroom observation ratings across years, including the percent of Hispanic students, percent of Black students, percent of males, percent of students with disabilities, and percent of English language learners (ELLs). Therefore, this dissertation used the percent of minority students, percent of male students, percent of ELLs, percent of students with disabilities, percent of students eligible for free or reduced lunch, and class size as the classroom-level predictors in the analyses.

Review of the Cross-Classified Random Effects Model (CCREM)

The CCREM is an extension of typical multilevel model to analyze data with cross-classification structures. In this dissertation, a CCREM was utilized to examine the classroom variance components and how differential classroom characteristics contributed to the variation in the lower-level ratings (i.e., multiple ratings nested within classrooms). This section introduces multilevel modeling and cross-classified random effects modeling using equations and examples.

Introduction of multilevel modeling

Multilevel modeling is a statistical method to analyze data with hierarchical structures (e.g., students nested within classrooms within schools) that are common in a variety of applications, including studies of growth and organizational effects (Raudenbush & Bryk, 2002). In educational settings, hierarchical data structures are seen frequently, for example, students nested within classrooms within schools. If the nested data structure is not considered in the analysis, the assumption of independence of standard regression analysis will be violated (Raudenbush & Bryk, 2002). As stated by O’Connell and Reed (2012), “for clustered data, observations obtained from persons within the same cluster tend to exhibit more similarity to each other than to observations from different clusters” (p. 7). For example, if the gender gap in a student learning outcome is investigated using student achievement scores from multiple schools, ignoring school differences may generate biased results because the gender gap could vary across schools. Therefore, it is important to consider information from all levels of the analysis (Steenbergen & Jones, 2002). Additionally, multilevel modeling can estimate variance and covariance components with unbalanced, nested data (Raudenbush & Bryk, 2002). Moreover, multilevel modeling can help to examine how differential characteristics in the higher-level contexts contribute to explain the variation in lower-level outcomes (O'Connell & Reed, 2012). For example, the variation in lower-level outcomes (e.g., classroom observation ratings) may be impacted by the differences among higher-level groups or contexts (e.g., class size, teachers' year of experience).

As an illustration, to model observation ratings given to classrooms taught by teachers, correspondingly, the data would have a three-level hierarchical structure as seen in Figure 2.

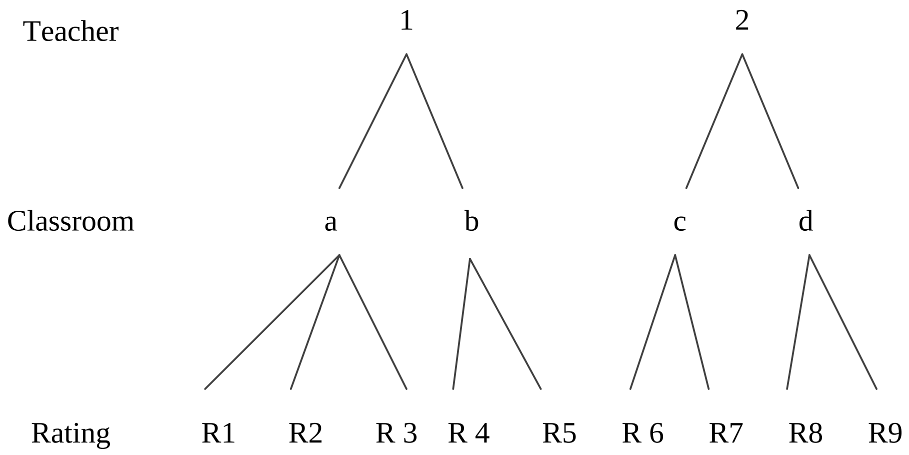

Figure 2. Network graph depicting three-level clustering of classroom observation ratings within classrooms within teachers.

In Figure 2, the level-1 units are observation ratings that are given to the level-2 units of classrooms nested within the level-3 units of teachers. In this case, Y_{ijk} is the score of rating (*i*) for classroom (*j*) taught by teacher (*k*). The formation at level 1 is

$$Y_{ijk} = \pi_{0jk} + e_{ijk}, \tag{1}$$

where π_{0jk} is the level-1 intercept, the mean rating of classroom *j* taught by teacher *k*, which is assumed to vary randomly at level 2. e_{ijk} is the level-1 residual, which is the deviation of the score Y_{ijk} from the classroom *jk*'s mean. e_{ijk} is a random "student effect", which is assumed normally distributed with a mean of zero and a constant level-1 variance, σ^2. The level-2 model for classrooms is

$$\pi_{0jk} = \beta_{00k} + r_{0jk}, \tag{2}$$

where β_{00k} is the level-2 intercept, the mean rating across classrooms taught by teacher *k*, which is assumed to vary randomly at level 3. r_{0jk} is the level-2 residual, which is the deviation of classroom *jk*'s mean from the teacher *k*'s mean. r_{0jk} is a random "classroom effect", which is assumed normally distributed with a mean of zero and a constant level-2 variance, $\tau_{\pi 00}$. The level-3 model for teachers is

$$\beta_{00k} = \gamma_{000} + u_{00k}, \tag{3}$$

where γ_{000} is the level-3 intercept, the grand mean. u_{00k} is the level-3 residual, which is the deviation of teacher *k*'s mean from the grand mean. u_{00k} is a random "teacher effect", which is assumed normally distributed with a mean of zero and a constant level-3 variance, $\tau_{\beta 00}$. The single equation for the three-level model is

$$Y_{ijk} = \gamma_{000} + u_{00k} + r_{0jk} + e_{ijk}. \tag{4}$$

This model provides information about the variation of classroom observation ratings at each of the three levels. σ^2 refers to the variation of ratings within classrooms within teachers. $\tau_{\pi 00}$ refers to the variation of ratings among classrooms within teachers. $\tau_{\beta 00}$ refers to the variation of ratings among teachers. This is an unconditional model because there are no predictors included (Raudenbush & Bryk, 2002).

One useful index called the intraclass correlation coefficient (ICC) indicates the proportion of the variance in the outcome that is between units (Raudenbush & Bryk, 2002). For example, the proportion of the variance in the ratings between classrooms within teachers is

$$ICC_{jk} = \frac{\tau_{\pi 00}}{\sigma^2 + \tau_{\pi 00} + \tau_{\beta 00}}. \qquad (5)$$

In the above unconditional three-level model, there are residuals (i.e., e_{ijk}, r_{0jk}, and u_{00k}), two random coefficients (i.e., π_{0jk} is level-1 random coefficient and β_{00k} is level-2 random coefficient), and the point estimate of the grand mean, γ_{000}. Multilevel modeling can also be estimated by adding predictors to each level. If we are interested in investigating whether the classroom observation ratings vary across classrooms due to the differences in class size, class size can be used as a predictor to examine the relationship between the classroom observation ratings and the class size. Using the above example with class size (C_{jk}) as the predictor for the classroom level, the first level of this model is formulated the same as Equation 1. At level 2, the model is

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}\,(C_{jk}) + r_{0jk}, \qquad (6)$$

where β_{00k} is the mean rating across the classrooms taught by teacher *k* when class size (C_{jk}) equals zero. β_{01k} is the expected change in rating within teacher *k* for each unit increase in class size (C_{jk}). r_{0jk} is the intercept residual for classroom *j* taught by teacher *k* when class size (C_{jk}) equals zero. r_{0jk} is assumed normally distributed with a mean of zero and a constant level-2 variance, $\tau_{\pi 00}$. $\tau_{\pi 00}$ is defined as the variance of the mean rating within the teacher units after including the level-2 predictor, class size (C_{jk}). The formulation for the third level of this model with the influence of class size (C_{jk}) assumed as fixed is

$$\begin{cases}\beta_{00k} = \gamma_{000} + u_{00k} \\ \beta_{01k} = \gamma_{010}\end{cases}. \tag{7}$$

γ_{000} is the overall mean rating across classrooms and teachers when class size (C_{jk}) equals zero. u_{00k} is the intercept residual for teacher k when class size (C_{jk}) equals zero. u_{00k} is assumed normally distributed with a mean of zero and a constant level-3 variance, $\tau_{\beta 00}$. $\tau_{\beta 00}$ is defined as the variance of the mean rating among the teacher units after including class size (C_{jk}). β_{01k} is the class size (C_{jk}) effect for teacher k, which we assume is constant for all teachers at γ_{010}, a fixed class size (C_{jk}) effect.

From the perspective of variance components, after adding a predictor at level 2, some changes may occur in the estimation of $\tau_{\pi 00}$, the classroom variance. At level 2 in Equation 6, each $\tau_{\pi 00}$ estimate is a conditional variance. That is, the level-2 residual, r_{0jk}, is a residual classroom effect unexplained by the level-2 predictor, class size (C_{jk}). Likewise, each $\tau_{\pi 00}$ estimated in Equation 2 of the unconditional model is an unconditional level-2 variance. Comparison of the conditional variance with the unconditional variance indicates a substantial reduction in variance once the classroom-level factors (i.e., class size in this model) are taken into account (Raudenbush & Bryk, 2002). The proportion of the variance explained by the class size (C_{jk}) as the level-2 predictor is

$$\text{Proportion variation explained in } \pi_{0jk} = \frac{\tau_{\pi 00}\text{ (unconditional)} - \tau_{\pi 00}\text{ (conditional)}}{\tau_{\pi 00}\text{ (unconditional)}}. \tag{8}$$

The proportion reduction in variance will increase as significant predictors enter the model (Raudenbush & Bryk, 2002). However, the variance may stay the same or increase slightly if a truly nonsignificant predictor is incorporated in the model under Maximum Likelihood (ML) estimation (Raudenbush & Bryk, 2002).

Another important issue of multilevel modeling is the centering of predictors, which refers to the choice of predictor location. In Equation 6, β_{00k} is defined as the predicted mean rating across classrooms taught by teacher *k* with a value of zero on class size (C_{jk}). If the value of zero on class size (C_{jk}) is not meaningful (i.e., class size usually ranges from 10 to 20), a proper choice of centering the class size (C_{jk}) will be required in order to ease the interpretation and estimation (Raudenbush & Bryk, 2002). There are two broad ways of centering predictors within the clustering level, grand-mean centering and group-mean centering. In the case of grand-mean centering, Equation 6 can be represented as

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} (C_{jk} - \bar{C}_{..}) + r_{0jk}, \quad (9)$$

where $\bar{C}_{..}$ refers to the class size mean averaged across all classrooms in the sample. The interpretation of β_{00k} is the mean rating across classrooms taught by teacher *k* when a classroom's class size equals the mean class size of all classrooms. β_{01k} is the expected change in the mean rating within teacher *k* for one unit increase in the adjusted class size (C_{jk}). In the case of group-mean centering, Equation 6 can be represented as

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} (C_{jk} - \bar{C}_{\cdot k}) + r_{0jk}, \quad (10)$$

where $\bar{C}_{\cdot k}$ refers to the class size mean averaged across the classrooms taught by teacher *k.* The interpretation of β_{00k} is the mean rating across classrooms taught by teacher *k* when a classroom's class size equals the mean class size of teacher *k*'s classrooms.

There are two broad estimation procedures to estimate the parameters of multilevel modeling, Maximum Likelihood (ML) estimation and Bayesian estimation (Field & Goldstein, 2006; Raudenbush & Bryk, 2002). ML estimation has been used across software (e.g., SAS software, HLM software) and research studies (see Hill & Goldstein, 1998; Rasbash & Goldstein, 1994). Bayesian estimation that uses a different language from ML estimation in describing point estimates, interval estimates, and hypothesis testing has been applied for multilevel data structures (see Browne & Draper, 2006; Field & Goldstein, 2006; Raudenbush & Bryk, 2002). ML estimation maximizes the joint likelihood of estimating the parameters (i.e., fixed effects and the variance/covariance components) for a fixed value of the sample data (Raudenbush & Bryk, 2002). There are differences between full ML estimation and restricted maximum likelihood (REML) estimation. REML estimation maximizes the joint likelihood of only the variance and covariance components given the observed sample data (Raudenbush & Bryk, 2002).

Introduction of cross-classified random effects modeling

With some advanced developments, multilevel modeling can be applied to cross-classified data structures for a variety of research purposes (Goldstein, 2003). In a typical hierarchical data structure, for example, classroom observation ratings nested within teachers, ratings only belong to a single element of a higher level. However, in reality, level-1 units are not clustered by one type of cluster and this type of purely nested data structure is not always found (Wallace, 2015). For instance, classroom observation ratings are clustered by teachers and by raters, while teachers may not be nested within raters. As a result, classroom observation ratings could be influenced by both teachers and raters. This type of data structure is called a cross-classified data structure.

In a classroom observation research paradigm, raters rate the teaching performance in the classrooms based on items from the observation protocol. As an illustration, if one particular teacher is rated by multiple raters at multiple occasions on each item of an observation instrument, correspondingly, ratings on each item are simultaneously nested within raters and teachers. In this case, for each item of the protocol, the data may appear as the example given in Table 1 and Figure 3.

Table 1

Cross-Classification Dataset Containing Classroom Observation Ratings Cross-Classified by Raters and Teachers

	Teacher		
Rater	1	2	3
A	R1, R2	R5, R6	
B	R3, R4		R9, R10
C		R7, R8	R11, R12

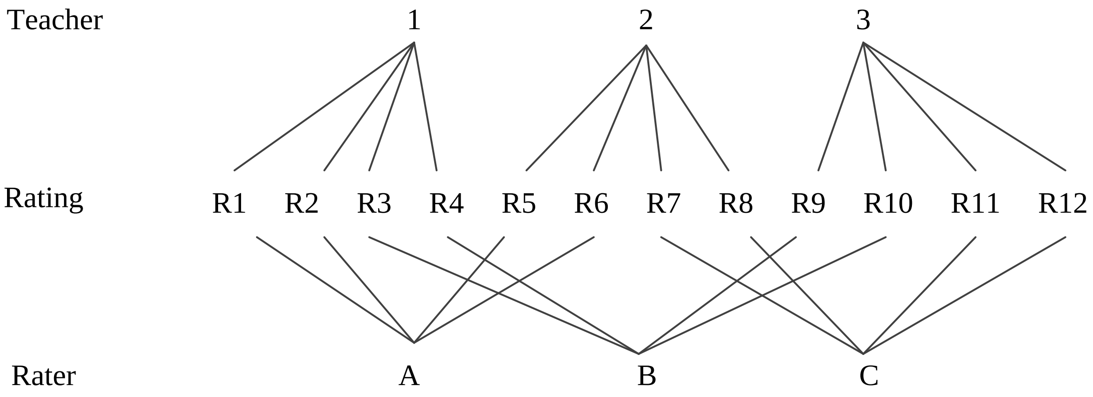

Figure 3. Network graph depicting clustering of ratings by teachers and cross-classification with raters.

The above example depicts a cross-classified data structure in which ratings are cross-classified by both teachers and raters. In this example, the rating $Y_{i\,(j_1,j_2)}$ on each classroom observation protocol item, given by rater j_1 to teacher j_2 can be modeled in a CCREM as

$$Y_{i\,(j_1,j_2)} = \beta_{0\,(j_1,j_2)} + e_{i\,(j_1,j_2)}, \tag{11}$$

and at level 2 (teachers and raters) as

$$\beta_{0\,(j_1,j_2)} = \gamma_{000} + u_{0j_10} + u_{00j_2} + u_{00j_1\times j_2}. \tag{12}$$

In Equations 11 and 12, the number of letters in the subscript represents the number of classifications (i.e., rating, rater, and teacher) (Rasbash & Browne, 2001). According to Beretvas (2001), the level-1 classification unit, rating, appears as the first subscript letter (i.e., "*i*") and the subscripts with the same common letter (i.e., "*j*") appearing in the parentheses separated by a comma represent the cross-classified factors (i.e., rater and teacher) at the same level.

In Equation 11, $\beta_{0\,(j_1,j_2)}$ is the mean rating in cell (j_1, j_2) (i.e., ratings given by rater j_1 to teacher j_2). $e_{i\,(j_1,j_2)}$ is the level-1 residual, which is the deviation of the rating $Y_{i\,(j_1,j_2)}$ from the predicted mean rating in cell (j_1, j_2). In Equation 12, γ_{000} is the grand mean rating. u_{0j_10} is the rater residual, which is the rater effect for rater j_1 averaged across teachers. u_{00j_2} is the teacher residual, which is the teacher effect for teacher j_2 averaged across raters. In Equations 11 and 12, three residuals, $e_{i\,(j_1,j_2)}$, u_{0j_10}, and u_{00j_2}, are assumed normally distributed with means of zero and their respective variances, σ^2_e, $\sigma^2_{u_{0j_10}}$, and $\sigma^2_{u_{00j_2}}$ (Beretvas, 2011). σ^2_e refers to the variation of ratings within the teacher by rater cell. $\sigma^2_{u_{0j_10}}$ and $\sigma^2_{u_{00j_2}}$ are between-rater variation and between-teacher variation in the ratings, respectively. In Equation 12, $u_{00j_1\times j_2}$ represents the random interaction effect between raters and teachers. This random interaction effect is usually set to zero because it is hard to separate its variance from the level-1 residual, σ^2_e, without sufficiently large within-cell sample sizes (Goldstein, 2003; Raudenbush & Bryk, 2002).

The intra-unit correlation coefficient (IUCC) of CCREMs functions similarly to the ICC in Equation 5 for the typical multilevel modeling, which represents the proportion of the variance in the outcome that is attributed to the units at each level (Raudenbush & Bryk, 2002). For instance, the IUCC at the teacher level represents the proportion of the variance shared by teachers, which can be calculated as

$$IUCC_{j_2} = \frac{\sigma^2_{u_{00j_2}}}{\sigma^2_e + \sigma^2_{u_{0j_10}} + \sigma^2_{u_{00j_2}}}. \tag{13}$$

If we are interested in using a teacher-level predictor, X_{j_1}, and a rater-level predictor, Z_{j_2}, to explain the variation in the intercept of ratings, Equation 12 becomes

$$\beta_{0\,(j_1,j_2)} = \gamma_{000} + \gamma_{010}(X_{j_1}) + \gamma_{020}(Z_{j_2}) + u_{0j_10} + u_{00j_2}. \tag{14}$$

In Equation 14, γ_{000} is the grand mean rating when X_{j_1} and Z_{j_2} equal zero. γ_{010} represents the expected change in the grand mean rating for one unit increase in X_{j_1} when Z_{j_2} equals zero. γ_{020} represents the expected change in the grand mean rating for one unit increase in Z_{j_2} when X_{j_1} equals zero.

The example above illustrates a cross-classified data structure with one level of cross-classification clustering. In a more complex case, in addition to the cross-classification of ratings by raters and teachers, the clustering of ratings within teachers' multiple classrooms may affect the ratings of interest (Beretvas, 2011). In this case, different raters rate each classroom at multiple occasions and some of these classrooms are taught by the same teacher. The data structure may appear as displayed in Table 2 and Figure 4. In Figure 4, there is a pure clustering of ratings within classrooms within teachers and there is a cross-classification of ratings by raters and classrooms within teachers. That means ratings are cross-classified by raters and classrooms, and ratings are also cross-classified by raters and teachers.

Table 2

Cross-Classification Dataset Containing Classroom Observation Ratings Cross-Classified by Raters and Classrooms Nested within Teachers

Rater	Teacher 1		Teacher 2	
	Class a	Class b	Class c	Class d
A	R1, R2	R5, R6		
B	R3, R4		R7, R8	
C			R9, R10	R11, R12

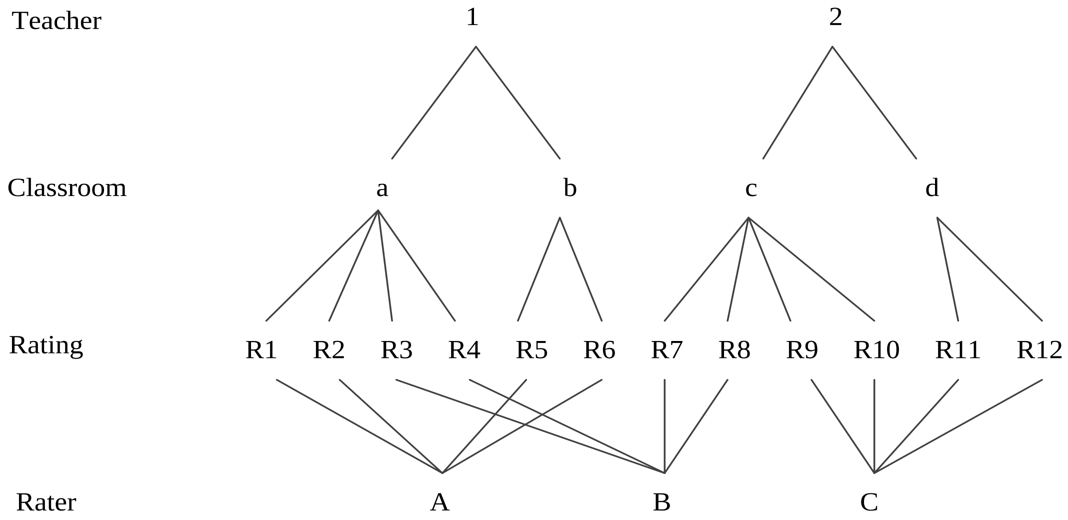

Figure 4. Network graph depicting clustering of ratings by classrooms within teachers and cross-classification with raters.

To model this data structure for the ratings on each item of a classroom observation instrument, $Y_{i\,(jk_1,k_2)}$ is the score of rating (*i*) for classroom (*j*) taught by teacher (k_1), which is given by raters (k_2). The unconditional model formation at level 1 is

$$Y_{i\,(jk_1,k_2)} = \pi_{0\,(jk_1,k_2)} + e_{i(jk_1,k_2)}, \tag{15}$$

and at level 2 (classrooms):

$$\pi_{0\,(jk_1,k_2)} = \beta_{00(k_1,k_2)} + u_{0jk_1}, \tag{16}$$

and at level 3 (teachers and raters):

$$\beta_{00(k_1,k_2)} = \gamma_{0000} + v_{000k_1} + \ v_{000k_2}, \tag{17}$$

and as a single equation:

$$Y_{i\,(jk_1,k_2)} = \gamma_{0000} + v_{000k_1} + v_{000k_2} + u_{0jk_1} + e_{i(jk_1,k_2)}. \tag{18}$$

In Equation 15, $\pi_{0\,(jk_1,k_2)}$ is the mean rating given by rater k_2 to classroom *j* taught by teacher k_1. $e_{i(jk_1,k_2)}$ is the level-1 residual, the deviation of the rating $Y_{i\,(jk_1,k_2)}$ from the mean rating in cell (jk_1, k_2). In Equation 16, $\beta_{00(k_1,k_2)}$ is the predicted mean rating given by rater k_2 averaged across the classrooms of teacher k_1. u_{0jk_1} is the level-2 residual, the deviation of mean rating in cell (jk_1, k_2) from the predicted mean rating in cell (k_1, k_2) averaged across classrooms. In Equation 17, γ_{0000} is the grand mean rating. v_{000k_1} is the teacher residual, which is the teacher effect for teacher k_1 averaged across raters. v_{000k_2} is the rater residual, which is the rater effect for rater k_2 averaged across teachers. In Equation 18, four variance components are associated with the four residuals, v_{000k_1}, v_{000k_2}, u_{0jk_1}, and $e_{i(jk_1,k_2)}$. Each residual is assumed normally distributed with a mean of zero and respective variances $\sigma^2_{v_{000k_1}}$ for v_{000k_1} of the teacher level, $\sigma^2_{v_{000k_2}}$ for v_{000k_2} of the rater level, $\sigma^2_{u_{0jk_1}}$ for u_{0jk_1} of the classroom level, and σ^2_e for $e_{i(jk_1,k_2)}$ of the level 1.

According to Murphy and Beretvas (2015), a CCREM is appropriate when there are multiple scores provided by multiple raters per item per teacher even when there is an unbalanced number of raters per item. Murphy and Beretvas (2015) compared the rater effects estimates using two scaling methods (i.e., the classical test theory and item response theory) and three models, including the conventional multilevel model, the CCREM, and the cross-classified multiple membership random effects model (CCMMrem). The results showed that ignoring rater effects could lead to teachers being misclassified, and better estimates of teacher effectiveness were produced using a CCREM regardless of the scaling method. Moreover, ignoring or misspecifying the cross-classification structure (i.e., modeling cross-classified data structure using a conventional multilevel model) may generate biased fixed effects estimates, standard error estimates, and variance component estimates (Fielding & Goldstein; 2006; Luo & Kwok, 2010; Meyers & Beretvas, 2006; Rasbash & Browne, 2001; Wallace, 2015).

Summary

Among teacher effectiveness measures, the classroom observation is an important component of most states' evaluation systems (Hull, 2013; National Council on Teacher Quality, 2015; Whitehurst et al., 2014). If a teacher's classroom observation ratings vary greatly across the classrooms he or she teaches, ratings from one single classroom cannot be representative for all of his or her classrooms. This dissertation examined the variation of a teacher's classroom observation ratings across his or her multiple classrooms as a function of a set of classroom-level predictors for the interpretation and use of classroom observation ratings in teacher evaluations. Ratings on each dimension of the CLASS instrument collected by the MET project were analyzed using a CCREM. The next section, Chapter 3, includes a description of the methods and procedures for the analyses of this dissertation.

3 METHODOLOGY

This dissertation used a cross-classified random effects model (CCREM) to examine the variation of classroom observation ratings across teachers' multiple classrooms as a function of classroom characteristics. Sample sizes, data structures, and statistical analysis procedures are presented in this chapter.

Data Sources and Sample

The researchers of the Measures of Effective Teaching (MET) project collected a variety of measures regarding teaching quality in classrooms over a two-year period (Academic Years 2009-2010 and 2010-2011) in the United States. More than 2,500 teachers in grades four through nine participated in the study (White & Rowan, 2013). These teachers worked for 317 different schools that were distributed throughout the following six large school districts: Charlotte-Mecklenburg (NC) Schools, Dallas (TX) Independent School District, Denver (CO) Public Schools, Hillsborough County (FL) Public Schools, Memphis (TN) City Schools, and New York City (NY) Department of Education (White & Rowan, 2013).

In Year 1 of the MET project (Academic Year 2009-2010), most of the specialist teachers (i.e., teachers who only taught a single subject, ELA or math) in grades six to nine and a handful of specialist teachers in grades four to five taught multiple classrooms of students (White & Rowan, 2013). The MET researchers collected classroom observation data from two classrooms by these teachers (White & Rowan, 2013). There were only two classrooms within each teacher, which is considered as a small within-group sample size. This, however, reflects a reality that teachers typically do not teach many classrooms at one period of time. In this dissertation, the number of groups was large (i.e., teacher sample sizes were 414 for math and 458 for ELA). Theall et al. (2011) found that the fixed and random effects parameter estimates were not affected by small within-group size (e.g., *n* = 2) for both unconditional and conditional models when the number of groups was large (e.g., *n* = 459). For the analysis of this dissertation, two datasets with teachers who had two classrooms in Year 1 were created for math and ELA content areas. Sample sizes at each level and within each level for the two datasets are displayed in Tables 3 and 4.

Table 3

Sample Sizes at Each Level

Units	Math	ELA
Ratings	3,850	4,264
Raters	260	256
Classrooms	828	915
Teachers	414	458
Schools	137	142

Table 4

Sample Sizes within Each Level

Subject	Nesting Structure	Min.	Max.	*M*
ELA	Ratings within raters	1	99	16.73
	Ratings within Schools	4	123	26.88
	Ratings within Teachers	3	15	8.33
	Ratings within Classrooms	1	8	4.17
	Classrooms within teachers	2	2	2.00
	Teachers within schools	1	13	3.23
Math	Ratings within raters	1	107	14.81
	Ratings within Schools	4	105	24.85
	Ratings within Teachers	4	12	8.22
	Ratings within Classrooms	1	8	4.11
	Classrooms within teachers	2	2	2.00
	Teachers within schools	1	10	3.02

Measures

The MET project applied multiple classroom observation protocols to measure the teaching quality in classrooms, including the Classroom Assessment Scoring System (CLASS), Framework for Teaching (FFT), Mathematical Quality of Instruction (MQI), Protocol for language Arts Teaching Observation (PLATO Prime), and Quality of Science Teaching (QST) (White & Rowan, 2013). The observation ratings of the CLASS instrument were used in this dissertation because it could be applied across multiple subjects (i.e., math and ELA). Another general instrument FFT was not chosen because only two of the four domains of the original FFT protocol were coded in the MET project.

The CLASS instrument is an observational protocol designed based on an extensive literature review on classroom practices and theories in human development and ecological systems to measure daily interactions between teachers and students across kindergarten through 12th grade (Pianta & Hamre, 2009). The CLASS has three broad domains of measurement (Emotional Support, Classroom Organization, and Instructional Support) with several dimensions belonging to each domain. Twelve models with ratings on each dimension as outcome variables were estimated for each of the subjects (i.e., math and ELA). Each dimension is an item measured according to specific behavioral indicators on a 7-point scale ranging from low to high (i.e., scores of 1 and 2 are considered to be in the low-range; 3, 4, and 5 are in the mid-range; and 6 and 7 are in the high-range).

Classroom observation ratings typically involve an ordinal scale, which may not satisfy the assumption of normality required in many statistical procedures (Murphy & Beretvas, 2015; Raudenbush & Bryk, 2002). In practice, classroom observation ratings are commonly averaged across ratings as if they were on an interval scale with a normal distribution (Murphy & Beretvas, 2015). However, the aggregation procedure will result in losing information from the data (Hox et al., 2010; Raudenbush & Bryk, 2002). In addition, some prior studies showed that a 7-point scale variable with an underlying measurement continuum could be assumed as a continuous variable (e.g., Carifil & Perla, 2007; Lubke & Muthén, 2004; Norman, 2010; Rhemtulla, Brosseau-Liard, & Savalei, 2012). Therefore, this dissertation analyzed the 7-point scale classroom observation ratings on each dimension (i.e., item) of the CLASS instrument as continuous outcome variables.

The information of the domains and dimensions of the CLASS instrument and the distributions of each dimension are displayed in Table 5. There are no agreed-upon cutoff values to judge if a variable is normally distributed based on Skewness and Kurtosis (Finney & DiStefano, 2006). However, for sample sizes greater than 300, it is recommended if the Skewness is larger than 2 or the Kurtosis is larger than 7, problems may occur under Maximum Likelihood (ML) estimation (West, Finch, & Curran, 1995). Therefore, most of the dimensions in Table 5 can be regarded as normally distributed continuous variables except the Negative climate dimension.

Table 5

Means, Standard Deviations, Skewness, and Kurtosis Values for the CLASS Dimensions

		ELA				Math			
Domain	Dimension	*M*	*SD*	Skewness	Kurtosis	*M*	*SD*	Skewness	Kurtosis
Emotional Support	Positive climate	4.33	1.30	−0.08	−0.57	4.10	1.29	0.02	−0.54
	Negative climate	1.44	0.80	2.49	8.26	1.50	0.83	2.17	6.10
	Teacher sensitivity	4.01	1.28	0.02	−0.49	4.05	1.23	0.01	−0.46
	Regard for student perspectives	3.20	1.33	0.33	−0.56	2.71	1.16	0.61	0.07
Classroom Organization	Behavior management	5.85	1.19	−1.31	1.74	5.76	1.25	−1.25	1.32
	Productivity	5.75	1.14	−1.15	1.49	5.63	1.20	−1.10	1.17
	Instructional learning formats	4.06	1.20	−0.18	−0.41	3.95	1.17	−0.05	−0.42
Instructional Support	Content understanding	3.65	1.28	0.02	−0.50	3.61	1.18	0.08	−0.37
	Analysis and problem solving	2.62	1.23	0.74	0.26	2.37	1.09	0.90	0.81
	Quality of feedback	3.63	1.28	0.11	−0.51	3.32	1.22	0.30	−0.26
	Instructional dialogue	3.17	1.36	0.37	−0.48	2.92	1.18	0.45	−0.22
Student Engagement	Student engagement	4.65	1.17	−0.30	−0.18	4.53	1.15	−0.20	−0.12

Classroom-Level Predictors

In order to investigate how classroom observation ratings vary as a function of classroom characteristics, related classroom-level predictors should be added to examine how these predictors explain the classroom-level variation in observation ratings. Classroom characteristics could be measured by contextual characteristics such as class size (Marsh et al., 2012) and compositional characteristics such as student composition (Dreeben & Barr, 1988; Hattie, 2002). Moreover, Polikoff (2015) used student characteristics in the classrooms as predictors to explain the variation of classroom observation ratings across years, such as percent of Hispanic students, percent of Black students, and percent of males. Therefore, to investigate the second research question of this dissertation, the classroom-level predictors were added to each model, including class size, percent of minority students, percent of male students, percent of English language learners (ELLs), percent of students with disabilities, and percent of students eligible for free or reduced lunch. If these classroom characteristics were significant predictors, the classroom-level variation in classroom observation ratings was expected to appear less. These classroom-level predictors and their coding are displayed in Table 6.

Table 6

Classroom-Level Predictors and Coding

Variable	Label	Coding
CSIZE	The number of students ever listed in the given classroom	Mean = 24
MALE	The percent of students in the classroom who are male	Percent = 0 to 1
ELL	The percent of students in the classroom who are English language learners	Percent = 0 to 1
DISABILITY	The percent of students in the classroom with disabilities	Percent = 0 to 1
FRL	The percent of students in the classroom eligible for free or reduced price lunch	Percent = 0 to 1
MINOR	The percent of minority students in the classroom	Percent = 0 to 1

Note. Minority students are students who are not White students.

Analytical Approach

According to White and Rowan (2013), selected teachers agreed to have their classroom instructions observed on several occasions during each school year of the MET project. The raters were trained between 17 and 25 hours by self-directed websites established by the protocol developers. Each rater scored each classroom he or she observed in an online system based on the twelve dimensions of the CLASS instrument. In the datasets created for this dissertation, each classroom had one to four videos recorded from different times, where each video was divided by two 15-minute segments. The raters scored all the dimensions (i.e., items) of the CLASS instrument on these segments. Therefore, in the datasets, each classroom had multiple ratings on the segments from multiple videos for each item of the CLASS instrument.

Ideally, the segment ratings within videos and the videos within classrooms would be modeled as two levels statistically. The MET data structure would have ratings on segments within videos[1], videos within classrooms, classrooms within teachers, teachers within schools that are cross-classified by raters. However, this dissertation did not model the segment variance within videos and the video variance within classrooms. Instead, it was assumed that the segments of videos within each classroom were interchangeable for several reasons. First, although it is desirable to model the segment variance within videos and the video variance within classrooms as two independent levels, a CCREM may not converge due to the inadequate sample size within each crossed “cell”. For example, in this dissertation, the number of segments within videos cross-classified by raters may not be enough for the model to converge because each video only has two segments. Second, although using the ratings averaged across the segments and videos on each dimension as the outcome variables is another possible option, using aggregated ratings would lose information (see Hox et al., 2010). Therefore, this dissertation did not model the segment variance within videos and the video variance within classrooms as independent levels. Instead, the structure of segment ratings cross-classified by raters and classrooms nested within teachers within schools was chosen as the model.

As an illustration of the cross-classified data structure used in this dissertation, Table 7 demonstrates an example of twenty-two ratings that are cross-classified corresponding to the level-1 ratings, the classroom level, the teacher level, the school level, and the rater level in Figure 5. Figure 5 shows two lower levels of clustering (i.e., classrooms within teachers) within one higher-level of clustering (i.e., schools). The crossing of the higher-level classifications results in the cross-classification factor (i.e., raters) being crossed with the lower-level clustering variables (i.e., classrooms and teachers). Therefore, ratings nested within classrooms within

[1] Six percent of the video segments were double rated by different raters. Sensitivity tests were conducted by deleting the double rated segments. Results generated from data without the double-rated segments were similar to the ones generated from the data with double rated segments. Therefore, this dissertation did not delete the double-rated segments.

teachers within schools are cross-classified by raters, where schools and raters are the two cross-classification factors. In this dissertation, this model was used to examine the classroom-level variation in classroom observation ratings as a function of classroom characteristics.

Table 7

Cross-Classification Dataset Containing Classroom Observation Ratings Cross-Classified by Raters and Classrooms Nested within Teachers within Schools

Rater	School 1				School 2			
	Teacher a		Teacher b		Teacher c		Teacher d	
	Class I	Class II	Class III	Class IV	Class V	Class VI	Class VII	Class VIII
A	R1, R2	R4, R5						
B	R3		R6, R7	R10, R11	R12			
C			R8, R9		R13, R14	R15, R16	R17, R18	R21, R22
D							R19, R20	

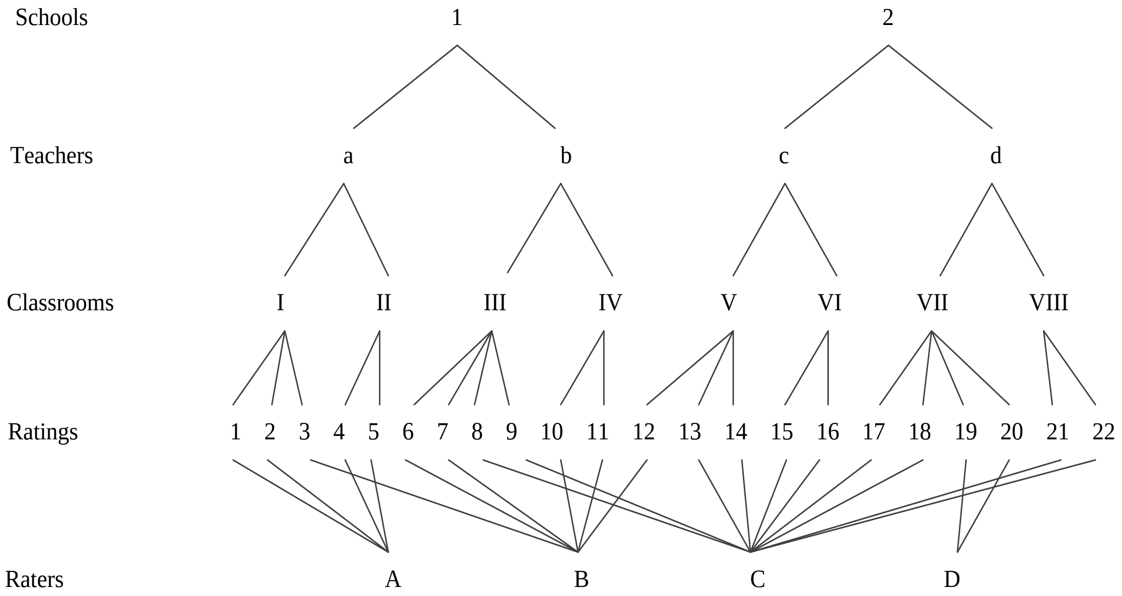

Figure 5. Network graph depicting clustering of ratings by classrooms within teachers within schools and cross-classification with raters.

R software (version 3.1.2; R Development Core Team, 2014) with the package lme4 (version 1.1-10; Bates, Maechler, Bolker, & Walker, 2015) was used to estimate the CCREMs. The lme4 package is an open sourced R package that can model cross-classified data at all levels using full ML or REML estimation. If the sample size at the clustering level is small, full ML may generate biased results in the estimation of variance components; otherwise, the two estimation procedures (i.e., full ML and REML) will produce very similar results (Raudenbush & Bryk, 2002). REML estimation provided as default by the lme4 package was used in this dissertation.

To answer the first research question, twelve models were estimated for the twelve dimensions (i.e., items) for each subject (i.e., math and ELA). The intra-unit correlation coefficient (IUCC) at the classroom level was used as an estimate of the proportion of the variance in classroom observation ratings shared by the classrooms. For each dimension of the

CLASS instrument, $Y_{i\,(jkl_1,l_2)}$ is the score of rating (*i*) of classroom (*j*) taught by teacher (*k*) in school (l_1) that was given by rater (l_2). The unconditional model formation at level 1 for each CLASS dimension is

$$Y_{i\,(jkl_1,l_2)} = \pi_{0\,(jkl_1,l_2)} + e_{i(jkl_1,l_2)}, \tag{19}$$

and at level 2 (classrooms):

$$\pi_{0\,(jkl_1,l_2)} = \beta_{00(kl_1,l_2)} + u_{0jkl_1}, \tag{20}$$

and at level 3 (teachers):

$$\beta_{00(kl_1,l_2)} = \gamma_{000(l_1,l_2)} + v_{00kl_1}, \tag{21}$$

and at level 4 (schools and raters):

$$\gamma_{000(l_1,l_2)} = \theta_{00000} + r_{0000l_1} + r_{0000l_2}, \tag{22}$$

and as a single equation:

$$Y_{i\,(jkl_1,l_2)} = \theta_{00000} + r_{0000l_1} + r_{0000l_2} + v_{00kl_1} + u_{0jkl_1} + e_{i(jkl_1,l_2)}. \tag{23}$$

In Equation 19, $\pi_{0\,(jkl_1,l_2)}$ is the mean classroom observation rating of rater l_2 given to classroom *j* taught by teacher *k* in school l_1. In Equation 20, $\beta_{00(kl_1,l_2)}$ represents the predicted mean rating given by rater l_2 averaged across the classrooms of teacher *k* in school l_1. In Equation 21, $\gamma_{000(l_1,l_2)}$ is the predicted mean rating given by rater l_2 averaged across the teachers in school l_1. In Equation 22, θ_{00000} is the grand mean rating. For the given dimension and subject in Equation 23, five variance components are associated with the five residuals, r_{0000l_1}, r_{0000l_2}, v_{00kl_1}, u_{0jkl_1}, and $e_{i(jkl_1,l_2)}$. Each residual is assumed normally distributed with a mean of zero and respective variances $\sigma^2_{r_{0000l_1}}$ for r_{0000l_1} of the school level, $\sigma^2_{r_{0000l_2}}$ for r_{0000l_2} of the rater level, $\sigma^2_{v_{00kl_1}}$ for v_{00kl_1} of the teacher level, $\sigma^2_{u_{0jkl_1}}$ for u_{0jkl_1} of the classroom level, and σ^2_e for $e_{i(jkl_1,l_2)}$ of the level 1. The IUCC at the classroom level was calculated as

$$IUCC_{0jkl_1} = \frac{\sigma^2_{u_{0jkl_1}}}{\sigma^2_{r_{0000l_1}} + \sigma^2_{r_{0000l_2}} + \sigma^2_{v_{00kl_1}} + \sigma^2_{u_{0jkl_1}} + \sigma^2_{e}}, \quad (24)$$

where IUCC_{0jkl_1} represented the proportion of the classroom variance in ratings within a particular teacher of the total variance.

To answer the second research question, the classroom-level predictors were added into the models, including the percent of minority students, percent of male students, percent of ELLs, percent of students with disabilities, percent of students eligible for free or reduced price lunch, and class size. All predictors were grand-mean centered across classrooms. The purpose was to explore to what extent these predictors could explain the classroom-level variation in the observation ratings. The formation for level 1 is the same as Equation 19. The formation for the classroom level is

$$\pi_{0\,(jkl_1,l_2)} = \beta_{00(kl_1,l_2)} + \beta_{01(kl_1,l_2)}(MINOR_{jkl_1} - \overline{MINOR_{\dots}}) + \beta_{02(kl_1,l_2)}(ELL_{jkl_1} - \overline{ELL_{\dots}}) + \cdots + \beta_{06(kl_1,l_2)}(DISABILITY_{jkl_1} - \overline{DISABILITY_{\dots}}) + u_{0jkl_1}, \quad (25)$$

and at level 3 (teachers):

$$\begin{cases} \beta_{00(kl_1,l_2)} = \gamma_{000(l_1,l_2)} + v_{00kl_1} \\ \beta_{01(kl_1,l_2)} = \gamma_{010(l_1,l_2)} \\ \beta_{02(kl_1,l_2)} = \gamma_{020(l_1,l_2)} \\ \vdots \\ \beta_{06(kl_1,l_2)} = \gamma_{060(l_1,l_2)} \end{cases}, \quad (26)$$

and at level 4 (schools and raters):

$$\begin{cases} \gamma_{000(l_1,l_2)} = \theta_{00000} + r_{0000l_1} + r_{0000l_2} \\ \gamma_{010(l_1,l_2)} = \theta_{01000} \\ \gamma_{020(l_1,l_2)} = \theta_{02000} \\ \vdots \\ \gamma_{060(l_1,l_2)} = \theta_{06000} \end{cases}, \tag{27}$$

and as a single equation:

$$Y_{i\,(jkl_1,l_2)} = \theta_{00000} + \theta_{01000}(MINOR_{jkl_1} - \overline{MINOR_{\ldots}}) + \theta_{02000}(ELL_{jkl_1} - \overline{ELL_{\ldots}}) + \cdots + \theta_{06000}(DISABILITY_{jkl_1} - \overline{DISABILITY_{\ldots}}) + r_{0000l_1} + r_{0000l_2} + v_{00kl_1} + u_{0jkl_1} + e_{i(jkl_1,l_2)}. \tag{28}$$

In Equation 25, $MINOR_{jkl_1}$, ELL_{jkl_1}, …, and $DISABILITY_{jkl_1}$represent the six classroom-level predictors (i.e., percent of minority students, percent of male students, percent of ELLs, percent of students with disabilities, percent of students eligible for free or reduced lunch, and class size) added to the models. $\beta_{00(kl_1,l_2)}$ is the predicted mean rating given by rater l_2 averaged across classrooms of teacher *k* in school l_1, when all the predictors equal to their means averaged across all the classrooms. $\beta_{01(kl_1,l_2)}$, $\beta_{02(kl_1,l_2)}$, …, and $\beta_{06(kl_1,l_2)}$ represent the expected changes in the mean rating given by rater l_2 for teacher *k* within school l_1 for one unit increase in each adjusted predictor when all the other five predictors equal to their means averaged across all the classrooms. In Equation 26, $\gamma_{000(l_1,l_2)}$ is the predicted mean rating given by rater l_2 averaged across teachers in school l_1, when all the predictors equal to their means averaged across all the classrooms. In Equation 27, θ_{00000} is the grand mean rating when all the predictors equal to their means averaged across all the classrooms. For the sake of simplicity, the influence of $MINOR_{jkl_1}$, ELL_{jkl_1}, …, and $DISABILITY_{jkl_1}$were estimated as fixed across teachers, schools, and raters. In Equation 28, θ_{01000}, θ_{02000}, …, and θ_{06000} represent the expected slopes for the classroom-level predictors.

For the random effects, five variance components are associated with the five residuals r_{0000l_1}, r_{0000l_2}, v_{00kl_1}, u_{0jkl_1}, and $e_{i(jkl_1,l_2)}$. Each residual is assumed normally distributed with a mean of zero and respective variances after including the classroom-level predictors $\sigma^2_{r_{0000l_1}}$ for r_{0000l_1} of the school level, $\sigma^2_{r_{0000l_2}}$ for r_{0000l_2} of the rater level, $\sigma^2_{v_{00kl_1}}$ for v_{00kl_1} of the teacher level, $\sigma^2_{u_{0jkl_1}}$ for u_{0jkl_1} of the classroom level, and σ^2_e for $e_{i(jkl_1,l_2)}$ of the level 1.

To answer the second question, the proportion of the variance explained by the classroom-level predictors at the classroom level were calculated as

$$\text{Proportion variation explained in } u_{0jkl_1} = \frac{\sigma^2_{u_{0jkl_1}}(\text{unconditional}) - \sigma^2_{u_{0jkl_1}}(\text{conditional})}{\sigma^2_{u_{0jkl_1}}(\text{unconditional})}. \qquad (29)$$

The proportion reduction in variance will increase as significant predictors enter the model (Raudenbush & Bryk, 2002). When there are multiple predictors entering the model, the proportion reduction in variance may jump to a higher value after the second significant predictor enters into the model. For example, the proportion reduction in variance by adding a predictor *X* to classroom-level can be calculated as

$$\text{Proportion variation explained by } X = \frac{\sigma^2_{u_{0jkl_1}}(\text{unconditional}) - \sigma^2_{u_{0jkl_1}}(X)}{\sigma^2_{u_{0jkl_1}}(\text{unconditional})}. \qquad (30)$$

Additionally, the proportion reduction in variance by adding the predictor *X* and a predictor *Z* to classroom-level can be calculated as

$$\text{Proportion variation explained by } X \text{ and } Z = \frac{\sigma^2_{u_{0jkl_1}}(\text{unconditional}) - \sigma^2_{u_{0jkl_1}}(XZ)}{\sigma^2_{u_{0jkl_1}}(\text{unconditional})}. \qquad (31)$$

Therefore, the incremental variance explained by adding *Z* to the model can be calculated as the difference between the proportion-variance-explained statistics in Equation 30 and Equation 31 (Raudenbush & Bryk, 2002). In this dissertation, each predictor's incremental variance was calculated to represent their contributions to explaining the classroom-level variation in the classroom observation ratings.

4 RESULTS

This section summarizes the results from the data analyses previously discussed in Chapter 3. This dissertation used a cross-classified random effects model (CCREM) to examine the variation of a teacher's classroom observation ratings across his or her multiple classrooms as a function of classroom characteristics. The first research question was examined by the unconditional models with each dimension (i.e., item) of the Classroom Assessment Scoring System (CLASS) instrument as dependent variables for math and ELA. The classroom-level intra-unit correlation coefficient (IUCC) indicates the proportion of the classroom variance in the classroom observation ratings. The second research question was examined by adding the classroom-level predictors (i.e., class size, percent of minority students, percent of male students, percent of English language learners (ELLs), percent of students with disabilities, and percent of students eligible for free or reduced lunch) to the models for math and ELA. The proportion reduction in classroom variance indicates how much the classroom characteristics explain the classroom-level variation in the classroom observation ratings. The results for the two research questions are presented in this chapter.

Classroom-Level Variation

Fixed effects estimates of the unconditional models

Table 8 provides the grand mean estimate from the math and ELA unconditional models across all the dimensions of the CLASS instrument.

Table 8

Fixed Effects of the Unconditional Models for Math and ELA across the Dimensions of the CLASS

		Math Dimension											
		PosC	NegC	Tsen	RgSP	BehM	PRD	ILF	ConU	APS	QuaF	InsD	Seng
Parameter	Coeff.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.
		(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)
Model for intercept													
Grand mean	θ_{00000}	4.176 (0.059)	1.554 (0.037)	4.093 (0.049)	2.775 (0.053)	5.668 (0.060)	5.530 (0.056)	3.941 (0.052)	3.691 (0.051)	2.458 (0.053)	3.415 (0.052)	3.022 (0.053)	4.562 (0.053)

		ELA Dimension											
		PosC	NegC	Tsen	RgSP	BehM	PRD	ILF	ConU	APS	QuaF	InsD	Seng
Parameter	Coeff.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.
		(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)
Model for intercept													
Grand mean	θ_{00000}	4.364 (0.056)	1.501 (0.035)	4.072 (0.051)	3.298 (0.054)	5.751 (0.057)	5.613 (0.053)	4.055 (0.051)	3.701 (0.055)	2.701 (0.053)	3.399 (0.054)	3.216 (0.052)	4.676 (0.051)

Note. Coeff. = coefficient; Est. = parameter estimate. Dimension abbreviations are as follows: PosC = Positive climate; NegC = Negative climate; Tsen = Teacher sensitivity; RgSP = Regard for student perspectives; BehM = Behavior management; PRD = Productivity; ILF = Instructional learning formats; ConU = Content understanding; APS = Analysis and problem solving; QuaF = Quality of feedback; InsD = Instructional Dialogue; Seng = Student engagement.

Variance component estimates

The IUCCs of all levels (i.e., classroom, teacher, school, rater, and residual) were calculated from the sources' variance divided by the sum of all the variance components. From Tables 9 to 20, the results of the variance component and standard deviation estimates for the unconditional and the conditional models of math and ELA are presented. Each table contains the results for one dimension of the CLASS instrument. For example, Table 9 provides the variance component and standard deviation estimates of the Positive climate dimension at each level for the unconditional and conditional models.

For the math unconditional models, the level-1 variance between ratings, $\sigma^2_{i(jkl_1,l_2)}$, varied from 0.382 to 0.999 depending on the dimension. The intercept variance between classrooms, $\sigma^2_{0jkl_1}$, varied from 0.067 to 0.227 depending on the dimension. Additionally, the intercept variance between teachers, $\sigma^2_{00kl_1}$, varied from 0.083 to 0.243 depending on the dimension. Further, the intercept variance between schools, $\sigma^2_{0000l_1}$, varied from 0.030 to 0.155 depending on the dimension. Finally, the intercept variance between raters, $\sigma^2_{0000l_2}$, varied from 0.118 to 0.421 depending on the dimension.

After the classroom-level predictors were added to the models, the variance component and standard deviation estimates might change due to the significance of the predictors. For the math conditional models, the level-1 variance between ratings, $\sigma^2_{i(jkl_1,l_2)}$, varied from 0.382 to 0.999 depending on the dimension. The intercept variance between classrooms, $\sigma^2_{0jkl_1}$, varied from 0.064 to 0.184 depending on the dimension. Moreover, the intercept variance between teachers, $\sigma^2_{00kl_1}$, varied from 0.085 to 0.262 depending on the dimension. Further, the intercept variance between schools, $\sigma^2_{0000l_1}$, varied from 0.015 to 0.105 depending on the dimension.

Finally, the intercept variance between raters, $\sigma^2_{0000l_2}$, varied from 0.116 to 0.417 depending on the dimension.

For the ELA unconditional models, the level-1 variance between ratings, $\sigma^2_{i(jkl_1,l_2)}$, varied from 0.368 to 1.215 depending on the dimension. The intercept variance between classrooms, $\sigma^2_{0jkl_1}$, varied from 0.080 to 0.227 depending on the dimension. The intercept variance between teachers, $\sigma^2_{00kl_1}$, varied from 0.080 to 0.228 depending on the dimension. Moreover, the intercept variance between schools, $\sigma^2_{0000l_1}$, varied from 0.041 to 0.148 depending on the dimension. The intercept variance between raters, $\sigma^2_{0000l_2}$, varied from 0.102 to 0.330 depending on the dimension.

For the ELA conditional models, the level-1 variance between ratings, $\sigma^2_{i(jkl_1,l_2)}$, varied from 0.369 to 1.216 depending on the dimension. Moreover, the intercept variance between classrooms, $\sigma^2_{0jkl_1}$, varied from 0.077 to 0.218 depending on the dimension. The intercept variance between teachers, $\sigma^2_{00kl_1}$, varied from 0.075 to 0.216 depending on the dimension. The intercept variance between schools, $\sigma^2_{0000l_1}$, varied from 0.027 to 0.113 depending on the dimension. Finally, the intercept variance between raters, $\sigma^2_{0000l_2}$, varied from 0.101 to 0.328 depending on the dimension.

Table 9

Random Effects Parameter and Standard Deviation Estimates of the Positive Climate Dimension for the Unconditional and Conditional Models of Math and ELA

		Model for Math			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.914	0.956	0.915	0.957
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.116	0.341	0.106	0.325
Teachers	$\sigma^2_{00kl_1}$	0.197	0.444	0.205	0.452
Schools	$\sigma^2_{0000l_1}$	0.080	0.283	0.068	0.260
Raters	$\sigma^2_{0000l_2}$	0.364	0.604	0.361	0.601
		Model for ELA			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.915	0.957	0.915	0.957
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.158	0.398	0.158	0.398
Teachers	$\sigma^2_{00kl_1}$	0.191	0.398	0.189	0.435
Schools	$\sigma^2_{0000l_1}$	0.094	0.307	0.064	0.253
Raters	$\sigma^2_{0000l_2}$	0.305	0.553	0.303	0.550

Note. Coeff. = coefficient; *Var*. = variance estimate.

Table 10

Random Effects Parameter and Standard Deviation Estimates of the Negative Climate Dimension for the Unconditional and Conditional Models of Math and ELA

		Model for Math			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.382	0.618	0.382	0.618
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.106	0.326	0.096	0.310
Teachers	$\sigma^2_{00kl_1}$	0.083	0.288	0.092	0.303
Schools	$\sigma^2_{0000l_1}$	0.030	0.174	0.015	0.121
Raters	$\sigma^2_{0000l_2}$	0.118	0.343	0.116	0.341
		Model for ELA			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.368	0.607	0.369	0.607
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.080	0.282	0.077	0.278
Teachers	$\sigma^2_{00kl_1}$	0.080	0.282	0.075	0.275
Schools	$\sigma^2_{0000l_1}$	0.041	0.201	0.027	0.165
Raters	$\sigma^2_{0000l_2}$	0.102	0.320	0.101	0.318

Note. Coeff. = coefficient; *Var*. = variance estimate.

Table 11

Random Effects Parameter and Standard Deviation Estimates of the Teacher Sensitivity Dimension for the Unconditional and Conditional Models of Math and ELA

		Model for Math			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.999	0.999	0.999	1.000
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.147	0.383	0.137	0.370
Teachers	$\sigma^2_{00kl_1}$	0.136	0.368	0.142	0.377
Schools	$\sigma^2_{0000l_1}$	0.047	0.218	0.043	0.206
Raters	$\sigma^2_{0000l_2}$	0.200	0.447	0.197	0.444

		Model for ELA			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	1.024	1.012	1.025	1.013
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.118	0.343	0.117	0.342
Teachers	$\sigma^2_{00kl_1}$	0.172	0.414	0.172	0.415
Schools	$\sigma^2_{0000l_1}$	0.067	0.259	0.052	0.228
Raters	$\sigma^2_{0000l_2}$	0.248	0.498	0.246	0.497

Note. Coeff. = coefficient; *Var*. = variance estimate.

Table 12

Random Effects Parameter and Standard Deviation Estimates of the Regard for Student Perspectives Dimension for the Unconditional and Conditional Models of Math and ELA

		Model for Math			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.779	0.883	0.779	0.883
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.119	0.344	0.111	0.333
Teachers	$\sigma^2_{00kl_1}$	0.096	0.310	0.099	0.315
Schools	$\sigma^2_{0000l_1}$	0.064	0.252	0.069	0.262
Raters	$\sigma^2_{0000l_2}$	0.312	0.559	0.314	0.560
		Model for ELA			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	1.148	1.071	1.149	1.072
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.141	0.375	0.135	0.368
Teachers	$\sigma^2_{00kl_1}$	0.151	0.388	0.157	0.396
Schools	$\sigma^2_{0000l_1}$	0.078	0.280	0.059	0.242
Raters	$\sigma^2_{0000l_2}$	0.275	0.525	0.274	0.524

Note. Coeff. = coefficient; *Var*. = variance estimate.

Table 13

Random Effects Parameter and Standard Deviation Estimates of the Behavior Management Dimension for the Unconditional and Conditional Models of Math and ELA

		Model for Math			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.762	0.873	0.762	0.873
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.200	0.447	0.184	0.429
Teachers	$\sigma^2_{00kl_1}$	0.243	0.493	0.262	0.512
Schools	$\sigma^2_{0000l_1}$	0.155	0.393	0.105	0.324
Raters	$\sigma^2_{0000l_2}$	0.205	0.453	0.204	0.452
		Model for ELA			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.634	0.796	0.634	0.796
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.227	0.477	0.218	0.467
Teachers	$\sigma^2_{00kl_1}$	0.228	0.477	0.216	0.465
Schools	$\sigma^2_{0000l_1}$	0.148	0.384	0.113	0.336
Raters	$\sigma^2_{0000l_2}$	0.223	0.472	0.222	0.471

Note. Coeff. = coefficient; *Var*. = variance estimate.

Table 14

Random Effects Parameter and Standard Deviation Estimates of the Productivity Dimension for the Unconditional and Conditional Models of Math and ELA

		Model for Math			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.771	0.878	0.771	0.878
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.110	0.331	0.106	0.326
Teachers	$\sigma^2_{00kl_1}$	0.133	0.365	0.136	0.369
Schools	$\sigma^2_{0000l_1}$	0.078	0.280	0.063	0.251
Raters	$\sigma^2_{0000l_2}$	0.355	0.596	0.353	0.594
		Model for ELA			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.711	0.843	0.711	0.843
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.088	0.297	0.085	0.292
Teachers	$\sigma^2_{00kl_1}$	0.085	0.292	0.088	0.297
Schools	$\sigma^2_{0000l_1}$	0.105	0.324	0.087	0.294
Raters	$\sigma^2_{0000l_2}$	0.316	0.563	0.316	0.563

Note. Coeff. = coefficient; *Var*. = variance estimate.

Table 15

Random Effects Parameter and Standard Deviation Estimates of the Instructional Learning Formats Dimension for the Unconditional and Conditional Models of Math and ELA

		Model for Math			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.847	0.920	0.847	0.920
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.074	0.272	0.072	0.268
Teachers	$\sigma^2_{00kl_1}$	0.135	0.368	0.136	0.369
Schools	$\sigma^2_{0000l_1}$	0.096	0.309	0.091	0.302
Raters	$\sigma^2_{0000l_2}$	0.227	0.476	0.226	0.475
		Model for ELA			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.876	0.936	0.876	0.936
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.106	0.325	0.104	0.322
Teachers	$\sigma^2_{00kl_1}$	0.132	0.363	0.134	0.367
Schools	$\sigma^2_{0000l_1}$	0.097	0.311	0.078	0.279
Raters	$\sigma^2_{0000l_2}$	0.225	0.475	0.226	0.475

Note. Coeff. = coefficient; *Var*. = variance estimate.

Table 16

Random Effects Parameter and Standard Deviation Estimates of the Content Understanding Dimension for the Unconditional and Conditional Models of Math and ELA

		Model for Math			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.895	0.946	0.896	0.946
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.070	0.265	0.072	0.268
Teachers	$\sigma^2_{00kl_1}$	0.128	0.358	0.125	0.354
Schools	$\sigma^2_{0000l_1}$	0.063	0.252	0.056	0.237
Raters	$\sigma^2_{0000l_2}$	0.265	0.515	0.264	0.514
		Model for ELA			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.981	0.990	0.981	0.991
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.108	0.329	0.109	0.330
Teachers	$\sigma^2_{00kl_1}$	0.149	0.386	0.150	0.387
Schools	$\sigma^2_{0000l_1}$	0.106	0.325	0.085	0.292
Raters	$\sigma^2_{0000l_2}$	0.280	0.529	0.279	0.528

Note. Coeff. = coefficient; *Var*. = variance estimate.

Table 17

Random Effects Parameter and Standard Deviation Estimates of the Analysis and Problem Solving Dimension for the Unconditional and Conditional Models of Math and ELA

		Model for Math			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.681	0.825	0.681	0.825
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.067	0.258	0.064	0.254
Teachers	$\sigma^2_{00kl_1}$	0.083	0.288	0.085	0.291
Schools	$\sigma^2_{0000l_1}$	0.034	0.185	0.033	0.183
Raters	$\sigma^2_{0000l_2}$	0.421	0.649	0.417	0.646
		Model for ELA			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.924	0.961	0.925	0.962
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.142	0.377	0.141	0.376
Teachers	$\sigma^2_{00kl_1}$	0.084	0.291	0.087	0.294
Schools	$\sigma^2_{0000l_1}$	0.069	0.264	0.056	0.236
Raters	$\sigma^2_{0000l_2}$	0.330	0.574	0.328	0.573

Note. Coeff. = coefficient; *Var*. = variance estimate.

Table 18

Random Effects Parameter and Standard Deviation Estimates of the Quality of Feedback Dimension for the Unconditional and Conditional Models of Math and ELA

		Model for Math			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.961	0.980	0.961	0.980
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.080	0.282	0.076	0.276
Teachers	$\sigma^2_{00kl_1}$	0.139	0.373	0.142	0.377
Schools	$\sigma^2_{0000l_1}$	0.068	0.262	0.065	0.256
Raters	$\sigma^2_{0000l_2}$	0.262	0.512	0.261	0.511
		Model for ELA			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	1.136	1.066	1.136	1.066
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.130	0.360	0.133	0.365
Teachers	$\sigma^2_{00kl_1}$	0.162	0.402	0.158	0.398
Schools	$\sigma^2_{0000l_1}$	0.105	0.324	0.087	0.295
Raters	$\sigma^2_{0000l_2}$	0.221	0.511	0.220	0.469

Note. Coeff. = coefficient; *Var*. = variance estimate.

Table 19

Random Effects Parameter and Standard Deviation Estimates of the Instructional Dialogue Dimension for the Unconditional and Conditional Models of Math and ELA

		Model for Math			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.832	0.912	0.832	0.912
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.080	0.283	0.077	0.278
Teachers	$\sigma^2_{00kl_1}$	0.095	0.308	0.096	0.310
Schools	$\sigma^2_{0000l_1}$	0.064	0.252	0.060	0.245
Raters	$\sigma^2_{0000l_2}$	0.331	0.576	0.331	0.575
		Model for ELA			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	1.215	1.102	1.216	1.103
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.165	0.406	0.162	0.402
Teachers	$\sigma^2_{00kl_1}$	0.134	0.367	0.135	0.368
Schools	$\sigma^2_{0000l_1}$	0.065	0.255	0.049	0.221
Raters	$\sigma^2_{0000l_2}$	0.261	0.509	0.260	0.510

Note. Coeff. = coefficient; *Var*. = variance estimate.

Table 20

Random Effects Parameter and Standard Deviation Estimates of the Student Engagement Dimension for the Unconditional and Conditional Models of Math and ELA

		Model for Math			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.762	0.873	0.761	0.873
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.093	0.304	0.085	0.292
Teachers	$\sigma^2_{00kl_1}$	0.109	0.331	0.117	0.342
Schools	$\sigma^2_{0000l_1}$	0.082	0.286	0.066	0.257
Raters	$\sigma^2_{0000l_2}$	0.281	0.530	0.281	0.530

		Model for ELA			
		Unconditional		Conditional	
Parameter	Coeff.	*Var.*	*SD*	*Var.*	*SD*
Level-1 variance between					
Ratings	$\sigma^2_{i(jkl_1,l_2)}$	0.784	0.885	0.784	0.885
Intercept variance between					
Classrooms	$\sigma^2_{0jkl_1}$	0.133	0.364	0.124	0.353
Teachers	$\sigma^2_{00kl_1}$	0.085	0.292	0.088	0.296
Schools	$\sigma^2_{0000l_1}$	0.077	0.278	0.055	0.234
Raters	$\sigma^2_{0000l_2}$	0.292	0.540	0.293	0.541

Note. Coeff. = coefficient; *Var*. = variance estimate.

Intra-unit correlation coefficient (IUCC)

This dissertation estimated five sources of variation that came from the facets of classroom, teacher, school, rater, and residual. The proportion of the variance at each level for the unconditional models was reported as the IUCCs, which were calculated by the sources' variance divided by the sum of all the variance components. In this dissertation, the residual variance included the segment-level residual variance, the video-level residual variance, the measurement error, many of the interactions, and all the unexplained errors. Table 21 provides the results of the IUCCs at all levels for the math and ELA unconditional models across the dimensions of the CLASS instrument.

Table 21

IUCCs at All Levels for the Math and ELA Unconditional Models across the Dimensions

	Math				
Dimension	Residual	Classroom	Teacher	School	Rater
Positive climate	.547	.069	.118	.048	.218
Negative climate	.531	.147	.115	.042	.164
Teacher sensitivity	.653	.096	.089	.031	.131
Regard for student perspectives	.569	.087	.070	.047	.228
Behavior management	.487	.128	.155	.099	.131
Productivity	.529	.082	.091	.054	.244
Instructional learning formats	.614	.054	.098	.070	.165
Content understanding	.630	.049	.090	.044	.186
Analysis and problem solving	.530	.052	.065	.026	.327
Quality of feedback	.636	.053	.092	.045	.174
Instructional dialogue	.607	.058	.047	.047	.241
Student engagement	.575	.069	.082	.062	.212
	ELA				
Dimension	Residual	Classroom	Teacher	School	Rater
Positive climate	.550	.095	.115	.057	.183
Negative climate	.548	.119	.119	.061	.152
Teacher sensitivity	.629	.072	.106	.041	.152
Regard for student perspectives	.640	.079	.084	.044	.153
Behavior management	.434	.155	.156	.101	.153
Productivity	.545	.067	.065	.080	.242
Instructional learning formats	.610	.074	.092	.068	.157
Content understanding	.604	.067	.092	.065	.172
Analysis and problem solving	.597	.092	.054	.045	.213
Quality of feedback	.648	.074	.092	.060	.126
Instructional dialogue	.660	.090	.073	.035	.142
Student engagement	.572	.097	.062	.056	.213

Classroom-level IUCCs

For both the math and ELA unconditional models, the Behavior management and Negative climate dimensions had generally larger IUCCs at the classroom level than other dimensions. For the Positive climate dimension, classrooms shared 6.9% of the variance in ratings for math and 9.5% of the variance in ratings for ELA. For the Negative climate dimension, classrooms shared 14.7% of the variance in ratings for math and 11.9% of the variance in ratings for ELA. For the Teacher sensitivity dimension, classrooms shared 9.6% of the variance in ratings for math and 7.2% of the variance in ratings for ELA. For the Student perspective dimension, classrooms shared 8.7% of the variance in ratings for math and 7.9% of the variance in ratings for ELA. For the Behavior management dimension, classrooms shared 12.8% of the variance in ratings for math and 15.5% of the variance in ratings for ELA. For the Productivity dimension, classrooms shared 8.2% of the variance in ratings for math and 6.7% of the variance in ratings for ELA. For the Instructional learning formats dimension, classrooms shared 5.4% of the variance in ratings for math and 7.4% of the variance in ratings for ELA. For the Content understanding dimension, classrooms shared 4.9% of the variance in ratings for math and 6.7% of the variance in ratings for ELA. For the Analysis and problem solving dimension, classrooms shared 5.2% of the variance in ratings for math and 9.2% of the variance in ratings for ELA. For the Quality of feedback dimension, classrooms shared 5.3% of the variance in ratings for math and 7.4% of the variance in ratings for ELA. For the Instructional dialogue dimension, classrooms shared 5.8% of the variance in ratings for math and 9.0% of the variance in ratings for ELA. For the Student engagement dimension, classrooms shared 6.9% of the variance in ratings for math and 9.7% of the variance in ratings for ELA.

To help visualize the proportion of the variance in ratings shared by classrooms for math and ELA across the twelve dimensions, Figure 6 displays the IUCCs at the classroom level of each dimension for the math and ELA unconditional models. The pattern of results for the IUCCs at the classroom level for math and ELA were generally similar. However, most of the dimensions for ELA had slightly larger IUCCs at the classroom level than for math (i.e., 2.0% to

3.2% larger depending on the dimension), except for the Negative climate, Teacher sensitivity, Regard for student perspectives, and Productivity dimensions.

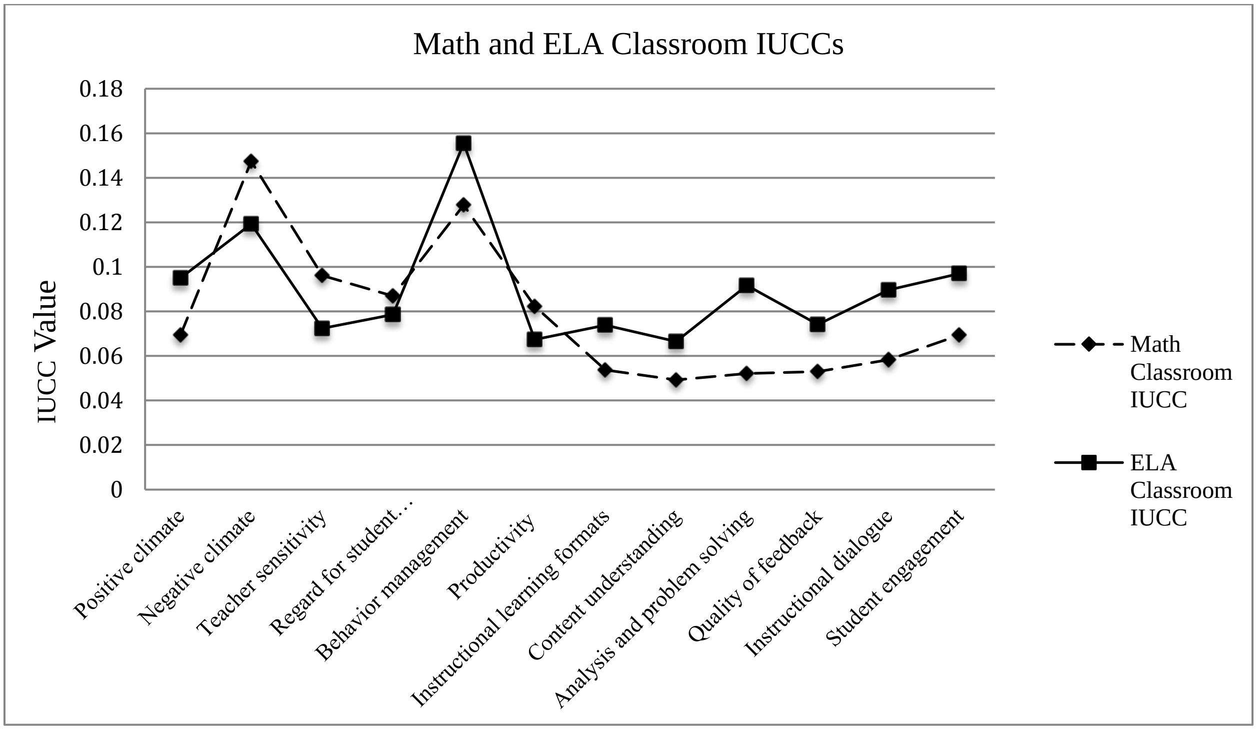

Figure 6. Proportion of variance explained at the classroom level for the math and ELA unconditional models across the dimensions.

IUCCs at all levels

As previously reported, the IUCCs at the classroom level revealed that the proportion of the variance in ratings explained by classrooms varied from 4.9% to 14.7% for math and 6.7% to 15.5% for ELA depending on the dimension. Table 20 also displays the proportion of variance in ratings explained by other facets of sources (i.e., teacher, school, rater, and residual). The proportion of the variance in ratings explained by teachers varied from 4.7% to 15.5% for math and 5.4% to 15.6% for ELA depending on the dimension. In addition, the proportion of variance in ratings explained by schools varied from 2.6% to 9.9% for math and 3.5% to 10.1% for ELA depending on the dimension. Moreover, the proportion of variance in ratings explained by raters varied from 13.1% to 32.7% for math and 12.6% to 24.2% for ELA depending on the dimension. Finally, the proportion of variance in ratings explained by residuals varied from 48.7% to 65.3% for math and 43.4% to 66.0% for ELA depending on the dimension.

To help visualize the proportion of variance in ratings explained at all levels for math and ELA across the twelve dimensions, Figure 7 displays the IUCCs at all levels across the dimensions for the math and ELA unconditional models. The patterns of the IUCCs for the math and ELA models were generally similar. However, for most of the dimensions, the IUCCs at the rater level for math were relatively higher than for ELA (i.e., .2% to 11.4% higher depending on the dimension), except for the Teacher sensitivity, Behavior management, and Student engagement dimensions. In addition, schools shared the least variance in ratings, while residuals shared the most variance in ratings.

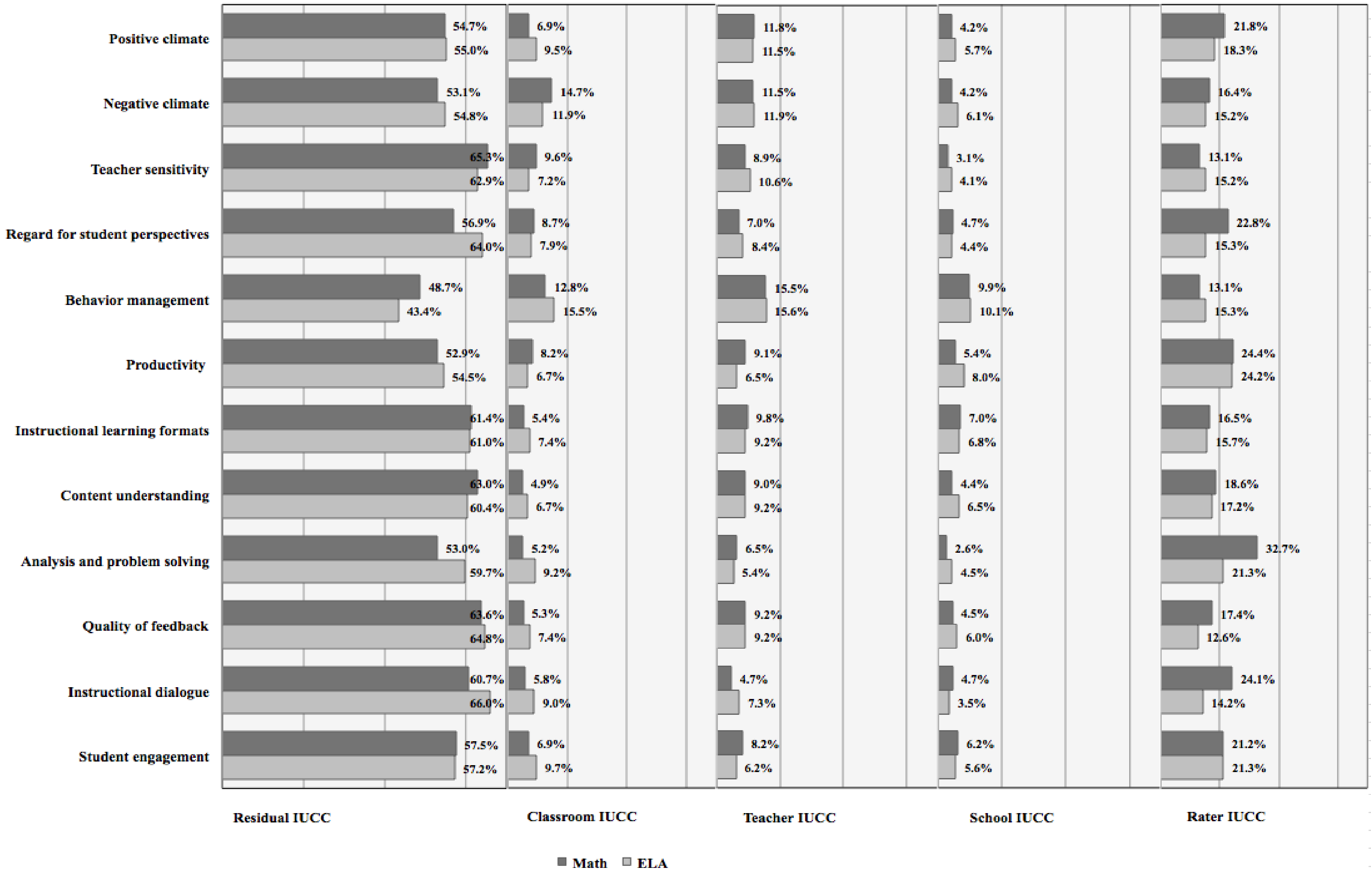

Figure 7. Proportion of variance explained at all levels for the math and ELA unconditional models across the dimensions. Bars represent the proportion of the variance in ratings at each level for the math and ELA unconditional models across the dimensions. Dark gray bars and light gray bars represent the variation at each level for the math and ELA models. Bars that cover the entire range would indicate a source of variation accounting for 70% of the variance in ratings.

Reduction in Variance at Classroom Level

As mentioned previously, the proportion reduction in variance will increase as significant predictors enter the model (Raudenbush & Bryk, 2002). However, the variance may stay the same or increase slightly if a truly non-significant predictor is incorporated in the model under the ML estimation in which zero or slightly negative variance reduction for the predictor may be computed (Raudenbush & Bryk, 2002). Table 22 displays the proportion reduction in variance at the classroom level by adding each predictor and all the classroom-level predictors (i.e., class size, percent of minority students, percent of ELLs, percent of male students, percent of students eligible for free or reduced lunch, and percent of students with disabilities) to the math and ELA models for all the dimensions of the CLASS instrument. After all the predictors were added, the variance in ratings at classroom level reduced −2.9% to 9.4% for the math models and −2.3% to 6.8% for the ELA models depending on the dimension. After these classroom-level predictors were added into both the math and ELA models, the classroom-level variation in the observation ratings on all the dimensions remained virtually unexplained. For the math models, the percent of minority students and the percent of ELLs contributed to explaining the classroom-level variation slightly depending on the dimension. Respectively for the ELA models, only the percent of minority students contributed to explaining the classroom-level variation slightly depending on the dimension.

Table 22

Incremental Proportion Reduction in the Classroom Variance by Adding Each Predictor for the Math and ELA Models

Dimension	Math						
	MINOR	*ELL*	*MALE*	*CSIZE*	*FRL*	*DISABILITY*	All Predictors
Positive climate	.069	.017	.017	−.008	−.009	.000	.086
Negative climate	.038	.037	.019	.000	.000	.000	.094
Teacher sensitivity	.054	.014	.007	−.007	−.007	−.007	.068
Regard for student perspectives	.025	.025	.009	.008	.000	.000	.067
Behavior management	.045	.015	.020	.000	.000	.000	.080
Productivity	.009	.046	.000	−.010	−.009	.000	.036
Instructional learning formats	.014	.040	−.013	.000	.000	−.014	.027
Content understanding	.000	−.015	.000	.000	−.014	.000	−.029
Analysis and problem solving	.030	.015	.000	.000	.000	.000	.045
Quality of feedback	.063	.013	−.013	.000	−.013	.000	.050
Instructional dialogue	.025	.013	.000	.000	−.013	.013	.038
Student engagement	.054	.032	.000	.000	.000	.000	.086
Dimension	ELA						
	MINOR	*ELL*	*MALE*	*CSIZE*	*FRL*	*DISABILITY*	All Predictors
Positive climate	.013	−.007	−.006	.006	.000	−.006	.000
Negative climate	.050	.000	−.012	.000	.000	.000	.038
Teacher sensitivity	.000	.008	.000	.009	−.009	.000	.008
Regard for student perspectives	.043	.007	−.015	−.007	−.007	.022	.043
Behavior management	.018	.000	.009	.018	−.005	.000	.040
Productivity	.011	.000	.000	.023	.000	.000	.034
Instructional learning formats	.028	.000	.000	.000	−.009	.000	.019
Content understanding	−.028	.009	−.009	.019	.000	.000	−.009
Analysis and problem solving	−.007	.007	.000	.000	.000	.007	.007
Quality of feedback	−.031	.000	−.007	.007	.000	.007	−.023
Instructional dialogue	.018	.012	−.006	−.006	−.006	.006	.018
Student engagement	.060	.000	.008	−.008	.008	.000	.068

Note. *MINOR* = percent of minority students; *ELL* = percent of ELLs; *MALE* = percent of male students; *CSIZE* = class size; *FRL* = percent of students eligible for free or reduced lunch; *DISABILITY* = percent of students with disabilities.

Furthermore, the results of the fixed effects parameter estimates for the math conditional models are presented in Table 23. There were relatively few statistically significant associations of the classroom characteristics with the observation ratings for math. Only the percent of minority students had a statistically significant negative relationship with most of the dimensions but a statistically significant positive relationship with the Negative climate dimension. The percent of ELLs and the percent of male students had statistically significant negative relationships with some of the dimensions but statistically significant positive relationships with the Negative climate dimension.

After the predictors were added in the math models, the impact of the percent of minority students, θ_{06000}, achieved statistical significance on most of the dimensions except for the Regard for student perspectives, Analysis and problem solving, and Instructional dialogue dimensions. That is, if the percent of minority students in the classroom increased, the adjusted means of most dimensions would decrease, while the adjusted mean of the Negative climate dimension would increase. In addition, the impact of the percent of ELLs, θ_{03000}, achieved statistical significance on the Negative climate, Productivity, and Quality of feedback dimensions. That is, if the percent of ELLs in the classroom increased, the adjusted means of the Productivity dimension and Quality of feedback dimension would decrease, while the adjusted mean of the Negative climate dimension would increase. Furthermore, the impact of the percent of male students, θ_{02000}, achieved statistical significance on the Negative climate and Behavior management dimensions. That is, if the percent of male students in the classroom increased, the adjusted mean of the Behavior management dimension would decrease, while the adjusted mean of the Negative climate dimension would increase.

Table 23

Fixed Effects Estimates for the Conditional Models for Math across the Dimensions of the CLASS

Parameter	Coeff.	Dimension											
		PosC	NegC	Tsen	RgSP	BehM	PRD	ILF	ConU	APS	QuaF	InsD	Seng
		Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.	Est.
		(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)	(*SE*)
Model for intercept													
Grand mean	θ_{00000}	4.127	1.643	3.884	2.931	5.599	5.521	3.958	3.854	2.549	3.445	3.251	4.614
		(0.214)	(0.155)	(0.217)	(0.198)	(0.229)	(0.200)	(0.196)	(0.157)	(0.174)	(0.203)	(0.191)	(0.154)
Model for slope													
CSIZE	θ_{01000}	0.000	0.000	00.001	0.005	−0.005	−0.004	−0.003	0.001	0.000	0.000	0.002	00.001
		(0.004)	(0.002)	(0.004)	(0.004)	(0.005)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
MALE	θ_{02000}	−0.223	0.378**	−0.126	−0.273	−0.451*	0.020	−0.062	0.122	−0.110	0.031	−0.073	−0.164
		(0.197)	(0.145)	(0.200)	(0.179)	(0.213)	(0.182)	(0.178)	(0.178)	(0.157)	(0.185)	(0.173)	(0.174)
ELL	θ_{03000}	−0.327	0.325*	−0.365	−0.231	−0.387	−0.451*	−0.306	−0.140	−0.187	−0.366*	−0.176	−0.332
		(0.198)	(0.140)	(0.194)	(0.178)	(0.217)	(0.181)	(0.181)	(0.177)	(0.153)	(0.185)	(0.171)	(0.174)
DISABILITY	θ_{04000}	0.011	−0.020	0.034	−0.015	0.002	0.012	0.001	−0.019	−0.017	−0.005	−0.031	−0.012
		(0.039)	(0.028)	(0.040)	(0.036)	(0.041)	(0.036)	(0.035)	(0.035)	(0.031)	(0.037)	(0.035)	(0.034)
FRL	θ_{05000}	0.000	0.000	0.001	−0.001	0.001	−0.001	−0.000	−0.001	0.000	0.000	−0.001	0.000
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
MINOR	θ_{06000}	−0.472**	0.367***	−0.404**	−0.090	−0.696***	−0.281*	−0.298*	−0.267*	−0.188	−0.260*	−0.216	−0.435***
		(0.141)	(0.095)	(0.133)	(0.128)	(0.158)	(0.130)	(0.133)	(0.125)	(0.107)	(0.132)	(0.122)	(0.125)

Note. Coeff. = coefficient; Est. = parameter estimate; *MINOR* = percent of minority students; *ELL* = percent of ELLs; *MALE* = percent of male students; *CSIZE* = class size; *FRL* = percent of students eligible for free or reduced lunch; *DISABILITY* = percent of students with disabilities. Dimension abbreviations are as follows: PosC = Positive climate; NegC = Negative climate; Tsen = Teacher sensitivity; RgSP = Regard for student perspectives; BehM = Behavior management; PRD = Productivity; ILF = Instructional learning formats; ConU = Content understanding; APS = Analysis and problem solving; QuaF = Quality of feedback; InsD = Instructional Dialogue; Seng = Student engagement. *p < .05. **p < .01. ***p < .001.

The results of the fixed effects parameter estimates for the ELA conditional models across all the dimensions of the CLASS instrument are presented in Table 24. There are relatively few statistically significant associations of the classroom characteristics with the observation ratings for ELA. Only the percent of minority students had a statistically significant negative relationship with most of the dimensions but a statistically significant positive relationship with the Negative climate dimension. The percent of male students and the class size had statistically significant negative relationships with some of the dimensions but statistically significant positive relationships with the Negative climate dimension.

After the predictors were added in the ELA models, the impact of the percent of minority students, θ_{06000}, achieved statistical significance on most of the dimensions except the Analysis and problem solving dimension. That is, if the percent of minority students in the classroom increased, the adjusted means of most dimensions would decrease, while the adjusted mean of Negative climate dimension would increase. In addition, the impact of the percent of male students, θ_{02000}, achieved statistical significance on the Positive climate, Negative climate, Teacher sensitivity, Behavior management, Productivity, and Student engagement dimensions. That is, if the percent of male students in the classroom increased, the adjusted means of the Positive climate, Teacher sensitivity, Behavior management, Productivity, and Student engagement dimensions would decrease, while the adjusted mean of the Negative climate dimension would increase. Furthermore, the impact of class size, θ_{01000}, achieved statistical significance on the Negative climate, Behavior management, and Productivity dimensions. That is, if the class size in the classroom increased, the adjusted means of the Behavior management dimension and Productivity dimension would decrease, while the adjusted mean of the Negative climate dimension would increase.

Table 24

Fixed Effects Estimates for the Conditional Models for ELA across the Dimensions of the CLASS

		Dimension											
		PosC	NegC	Tsen	RgSP	BehM	PRD	ILF	ConU	APS	QuaF	InsD	Seng
Parameter	Coeff.	Est. (*SE*)	Est. (*SE*)	Est. (*SE*)	Est. (*SE*)	Est. (*SE*)	Est. (*SE*)	Est. (*SE*)	Est. (*SE*)	Est. (*SE*)	Est. (*SE*)	Est. (*SE*)	Est. (*SE*)
Model for intercept													
Grand mean	θ_{00000}	4.342	1.495	4.061	3.501	5.570	5.580	4.168	3.868	2.907	3.510	3.331	4.663
		(0.175)	(0.114)	(0.169)	(0.177)	(0.177)	(0.146)	(0.159)	(0.167)	(0.164)	(0.178)	(0.181)	(0.154)
Model for slope													
CSIZE	θ_{01000}	−0.008	0.007*	−0.004	0.000	−0.017***	−0.009*	−0.001	−0.005	−0.005	−0.006	0.001	−0.005
		(0.005)	(0.003)	(0.005)	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)	(0.004)
MALE	θ_{02000}	−0.357*	0.448***	−0.339**	−0.186	−0.820***	−0.407**	−0.212	−0.306	−0.105	−0.255	−0.284	−0.495**
		(0.184)	(0.121)	(0.179)	(0.187)	(0.186)	(0.149)	(0.166)	(0.174)	(0.171)	(0.187)	(0.192)	(0.161)
ELL	θ_{03000}	0.103	−0.136	−0.181	−0.196	0.108	−0.100	−0.182	−0.230	−0.284	−0.146	−0.195	−0.098
		(0.214)	(0.140)	(0.208)	(0.215)	(0.220)	(0.175)	(0.195)	(0.204)	(0.194)	(0.218)	(0.218)	(0.184)
DISABILITY	θ_{04000}	0.005	−0.004	−0.003	−0.041	0.037	−0.010	−0.023	−0.029	−0.053	−0.032	−0.021	−0.017
		(0.037)	(0.025)	(0.036)	(0.038)	(0.038)	(0.031)	(0.034)	(0.036)	(0.035)	(0.038)	(0.039)	(0.033)
FRL	θ_{05000}	0.000	0.001	0.000	−0.001	0.000	0.000	0.000	−0.001	0.000	0.000	−0.001	0.001
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
MINOR	θ_{06000}	−0.591***	0.434***	−0.334**	−0.368**	−0.651***	−0.349**	−0.392**	−0.293*	−0.243	−0.378**	−0.376**	−0.564***
		(0.133)	(0.087)	(0.127)	(0.132)	(0.145)	(0.116)	(0.125)	(0.131)	(0.120)	(0.138)	(0.131)	(0.115)

Note. Coeff. = coefficient; Est. = parameter estimate. *MINOR* = percent of minority students; *ELL* = percent of ELLs; *MALE* = percent of male students; *CSIZE* = class size; *FRL* = percent of students eligible for free or reduced lunch; *DISABILITY* = percent of students with disabilities. Dimension abbreviations are as follows: PosC = Positive climate; NegC = Negative climate; Tsen = Teacher sensitivity; RgSP = Regard for student perspectives; BehM = Behavior management; PRD = Productivity; ILF = Instructional learning formats; ConU = Content understanding; APS = Analysis and problem solving; QuaF = Quality of feedback; InsD = Instructional Dialogue; Seng = Student engagement. **p* < .05. ***p* < .01. ****p* < .001.

5 DISCUSSION

For the interpretation and use of classroom observation ratings in teacher evaluations, this dissertation used a cross-classified random effects model (CCREM) to examine the variation of a teacher's classroom observation ratings across his or her multiple classrooms as a function of classroom characteristics. Two research questions were examined by a series of statistical analyses. The following section contains a discussion of the data analysis results presented in Chapter 4.

The Variation of Classroom Observation Ratings

Classroom-level variation in classroom observation ratings

Research question 1, what is the variation of a teacher's classroom observation ratings across his or her multiple classrooms, was examined by twelve models for math and ELA subjects. The intra-unit correlation coefficients (IUCCs) were calculated to show the proportion of variance in the classroom observation ratings explained by the classrooms, teachers, schools, raters, and residuals. Primarily, the results revealed that the variation of the observation ratings across teachers' multiple classrooms varied from 4.9% to 14.7% for math and 6.7% to 15.5% for ELA depending on the dimension (i.e., item) of the Classroom Assessment Scoring System (CLASS) instrument. That is, math classrooms accounted for 4.9% to 14.7% of the variance in ratings depending on the dimension. Additionally, ELA classrooms accounted for 6.7% to 15.5% of the variance in ratings depending on the dimension. This dissertation also found that the classroom-level variation in the classroom observation ratings could be as much as, or more than, the teacher-level variation (i.e., teachers accounted for 4.7% to 15.5% of the variance in ratings for math and 5.4% to 15.6% for ELA depending on the dimension). That is, classroom differences influenced the variation of classroom observation ratings as much as, or more than, teachers.

The results of this dissertation also demonstrated that the teaching quality measured by the CLASS instrument depended on both teachers and their classrooms. However, most of the prior research did not examine the variation of teachers' classroom observation ratings across their multiple classrooms; instead, ratings collected from one classroom per teacher represented both teacher and classroom effects (e.g., Bell et al., 2012; Hill et al., 2012; Smolkowski & Gunn, 2012). The proportion of variance explained by the classrooms in this dissertation showed that classrooms could be regarded as a source of variation in explaining classroom observation ratings. When classroom observation ratings are used to evaluate teachers in high-stakes settings, classroom-level variation should be considered as a construct-irrelevant variation.

Furthermore, the classroom-level variations in ratings on the Negative climate and Behavior management dimensions were relatively larger than other dimensions for both math and ELA. The Negative climate dimension measures the level of expressed negativity, such as anger, hostility, or aggression, demonstrated by teachers or students (The National Center on Quality Teaching and Learning, 2013). In a negative climate, teachers and students may not enjoy being together and spending time in the classroom (Stuhlman, Hamre, Downer, & Pianta, 2010). Moreover, the Behavior management dimension measures teachers' ability to use effective methods to prevent and redirect misbehavior by presenting clear behavioral expectations and minimizing time spent on behavioral issues (The National Center on Quality Teaching and Learning, 2013). The dimensions of the CLASS instrument were developed to measure the interactions between teachers and students (Pianta & Hamre, 2009). Based on what the Negative climate dimension and the Behavior management dimension measure regarding teaching quality, it is possible that students in the classroom influence the teaching quality under these two dimensions more than other dimensions.

Moreover, the classroom-level variation in the observation ratings was slightly larger for ELA than for math (i.e., 2.0% to 3.2% larger depending on the dimension) for most of the dimensions including the Positive climate, Behavior management, Instructional learning formats, Content understanding, Analysis and problem solving, Quality of feedback, Instructional dialogue, and Student engagement dimensions. Evertson, Anderson, Anderson, and Brophy (1980) found that there tended to be more interactions between students and teachers in ELA classrooms where students may be more active than in math classrooms. It is possible that students in ELA classrooms influence the teaching quality under some observational rubrics slightly more than students in math classrooms.

Rater-level variation in classroom observation ratings

The results revealed that raters were a large source of variation in the classroom observation ratings. This dissertation showed that raters accounted for 13.1% to 32.7% of the variance in ratings for math and 12.6% to 24.2% for ELA depending on the dimension. Bell et al. (2012) found that raters explained 5% to 30% of the variance in ratings depending on the CLASS domain. In addition, Kane and Staiger (2012) showed that raters explained 10% to 14% of the variance in ratings depending on the CLASS domain. Moreover, Hill et al. (2012) found that raters explained 4.56% to 28.58% of the variance in ratings depending on the Mathematical Quality of Instruction (MQI) dimension. These results indicate that some raters may be harsher or more lenient than other raters in evaluating teaching quality based on classroom observation instruments. Murphy and Beretvas (2015) suggested that teachers with lenient or severe raters would be more likely to be misclassified into a higher or lower performance category than they deserve. In order to minimize the rater effects, ongoing statistical monitoring of raters, for example, using a CCREM to account for rater bias or using multiple raters to score teaching quality in the classroom should be considered within the evaluation system (Murphy & Beretvas, 2015). Moreover, specific requirements for raters through hiring, training, and feedback should be focused to reduce the rater bias (Park, Chen, & Holtzman, 2014).

In addition, the rater-level variation in this dissertation was generally larger for math than ELA classrooms (i.e., .2% to 11.4% larger depending on the dimension) for most of the dimensions, including the Positive climate, Negative climate, Regard for student perspectives, Productivity, Instructional learning formats, Content understanding, Analysis and problem solving, Quality of feedback, and Instructional dialogue dimensions. That could be due to raters scoring ELA classrooms more consistently than math classrooms. Hill et al. (2012) suggested that it is important to examine if rater consistency varies depending upon diverse subjects (e.g., English and math). In this dissertation, the results suggested that rater consistency was generally lower for math classrooms than ELA classrooms for most of the CLASS dimensions. It is possible that raters for math classrooms need more rigorous rater selection and training to enhance the rater consistency.

School-level variation in classroom observation ratings

The results of this dissertation found that schools accounted for 2.6% to 9.9% of the variance in ratings for math and 3.5% to 10.1% for ELA depending on the dimension. Bell et al. (2012) suggested that school policy, school climate, and school leadership might contribute to teaching quality. In addition, ignoring a level of nesting in a multilevel analysis can impact the estimates of variance components and fixed effects, and the standard errors estimates of the coefficients of the lower level variables will generally be smaller resulting in inflated Type I error rates (Hox et al., 2010; Raudenbush & Bryk, 2002). These results indicate that school differences should be considered in examining the variation of classroom observation ratings when teachers from different schools are evaluated using observation ratings.

Residual variance in classroom observation ratings

In this dissertation, the majority of the variation in the observation ratings was attributed to the residual error (i.e., residual error shared 48.7% to 65.3% of the variance in ratings for math and 43.4% to 66.0% of the variance in ratings for ELA). The residual error included the segment residual variance, the video residual variance, the measurement error, many of the interactions, and all the unexplained errors. The segment residual variance and the video residual variance were regarded as parts of the residual component due to the segment variance within videos and the video variance within classrooms not being modeled independently. In addition, prior studies showed that residual error generally accounted for a large proportion of the variance in classroom observation ratings. For example, Kane and Staiger (2012) found that residual error accounted for 32% to 42% of the variance in ratings depending on the domain of the CLASS. Hill et al. (2012) found that residual error accounted for 32.97% to 44.77% of the variance in ratings depending on the dimension of the MQI instrument. Given that the variation in observation ratings due to the residual error is generally large, it is suggested that sampling should be well structured to capture multiple measurements within occasions (e.g., using multiple raters per classroom, capturing data from all segments of a class period) (Bell et al., 2012). It is also important to investigate the contributions of various aspects of factors that are sampled over to further investigate the teaching quality measured by classroom observations (Bell et al., 2012).

Classroom-Level Variation Explained by Classroom Characteristics

Research question 2, to what extent this variation of classroom observation ratings across teachers' multiple classrooms is explained by observable classroom characteristics, was examined by incorporating the classroom-level predictors (i.e., class size, percent of minority students, percent of male students, percent of English language learners (ELLs), percent of students with disabilities, and percent of students eligible for free or reduced lunch) into the math and ELA models. It is expected that different classrooms taught by the same teacher will receive different classroom observation ratings due to the differences in classroom demographic characteristics. However, the results suggested that the classroom-level predictors had limited contributions to explaining the classroom-level variation in the observation ratings. Only the percent of minority students contributed slightly to explaining the classroom-level variation for most of the dimensions. Moreover, the percent of ELLs, the percent of male students, and the class size contributed slightly to explaining the classroom-level variation for some of the dimensions depending on the subject. It is possible that there are other factors associated with classroom observation ratings that could explain the classroom-level variation, such as student belief and student knowledge (Bell et al., 2012).

Aligned with the reduction in variance results, there are relatively few statistically significant relationships between the classroom-level predictors and the classroom observation ratings. First, the percent of minority students had a statistically significant relationship with most of the dimensions. That is, classrooms with more minority students might receive lower ratings on most of the dimensions but higher ratings on the Negative climate dimension. Chaplin, Gill, Thompkins, and Miller (2014) found that teachers with more low-income and minority students tended to have lower observation ratings, while teachers with more gifted students tended to have higher ratings. It is possible that teachers have more difficulty in teaching classrooms with more ethnic or language minority students who come from diverse backgrounds (Reyhner, 1991; Rjosk, Richter, Hochweber, Lüdtke, & Stanat, 2015; Trueba, 1988). Moreover, the percent of ELLs, the percent of male students, and the class size had statistically significant relationships with some of the dimensions. That is, classrooms with more ELLs, male students, or a larger class size tended to receive lower ratings on some of the dimensions but higher ratings on the Negative climate dimension. It is possible that managing classrooms with more male students or a larger class size is more challenging for teachers (Blatchford et al., 2011; Fennema & Peterson, 1985). It also could be that raters tended to assign lower scores when they saw teachers leading classrooms with more minority students, ELLs, male students, or a larger class size under certain rubrics, regardless of teachers' performance (Whitehurst et al., 2014).

However, the magnitude of these predictors' effects on the classroom observation ratings should be considered when we interpret their practical importance (Kirk, 1996; Kirk, 2001). For example, the fixed effect estimates for the percent of minority students for the ELA classrooms varied between −0.243 and −0.651 depending on the dimension, which means that a 0.024 to 0.065 decrease in the rating will be expected with every additional 10% increase in the percent of minority students in ELA classrooms. The fixed effect estimates for the statistically significant predictors on a 7-point scale were relatively small from a practical perspective. Therefore, even if some of the predictors (i.e., percent of minority students, percent of male students, percent of ELLs, and class size) reached statistical significance, the fixed effect estimates for these predictors were virtually small.

6 CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH

Conclusions

This dissertation used a cross-classified random effects model (CCREM) to examine the classroom-level variation in the observation ratings and account for the teacher-level, school-level, and rater-level variation. The results demonstrated that teachers' multiple classrooms may receive different classroom observation ratings. Therefore, classroom-level variation should be taken into consideration when classroom observation ratings are used to evaluate teacher quality. In addition, a large proportion of the variance in ratings was attributed to the raters, schools, and residual error. These results suggested that one could model the observation ratings to control for these factors in teacher evaluations.

For the second research question, the classroom-level predictors (i.e., class size, percent of minority students, percent of male students, percent of English language learners (ELLs), percent of students eligible for free or reduced lunch, and percent of students with disabilities) had limited contributions to explaining the classroom-level variation in the classroom observation ratings. Moreover, some predictors (i.e., percent of minority students, percent of male students, percent of ELLs, and class size) were statistically significant but had practically small impacts on the ratings. Other classroom-level factors (e.g., student belief, student knowledge) that could contribute to explaining the classroom-level variation in classroom observation ratings should be investigated in future research.

Implications

The results of this dissertation indicate that teachers' multiple classrooms should be considered when classroom observation ratings are used to evaluate teachers in high-stakes settings. According to Bell et al. (2012), "the role of context is important for both professional development and human capital decisions; however, it is particularly important if observation scores are going to be used for high-stakes decisions regarding teachers" (p. 85). If teachers are awarded or denied based on the observation ratings that could be affected by contextual features outside of the teachers' control, high-stakes decisions may not be solid enough (Bell et al., 2012; Murphy & Beretvas, 2015). Because a teacher's observation ratings may vary across his or her multiple classrooms, using data from one classroom per teacher may give a distorted representation of teacher quality.

Additionally, when researchers and evaluators use observation ratings in practice, it should be clear which underlying construct of teacher effectiveness is used for inferences and decision-making. The purpose of the Classroom Assessment Scoring System (CLASS) instrument is to measure the interactions between teachers and students in the classroom (Pianta & Hamre, 2009). However, classroom observation ratings are sometimes used to make inferences and decisions regarding teacher performance only (e.g., the quality of instructional content or delivery) instead of the interactions between teachers and students. In this situation, in order to make inferences and decisions regarding teachers using classroom observation ratings, evaluators need to take the classroom context into consideration. Moreover, developing a systematic approach is needed for policy-makers and researchers to take the classroom-level variation into consideration when examining the variation of classroom observation ratings, such as collecting data from teachers' multiple classrooms and using a CCREM to account for the classroom context.

The majority of variation in the classroom observation ratings came from sources other than teachers and classrooms. The proportion of variance attributed to raters, schools, and residual error was the major source of variation in classroom observation ratings. Assessing how the variation in contexts (e.g., rater, classroom, school) affects the observation ratings can provide important information for identifying different contextual factors that influence the reported ratings (Hill et al., 2012). The results suggested that one could model the ratings to adjust for the effects of various contexts on the ratings because the observation systems may not provide a stable measure of teacher quality across these factors.

As recommended by Whitehurst et al. (2014), instead of using raw observation ratings to represent teacher quality, a systematic way of adjusting teacher observation ratings for certain student demographic characteristics should be developed to control for the classroom effects. This dissertation found that the classroom-level predictors (i.e., class size, percent of minority students, percent of male students, percent of ELLs, percent of students eligible for free or reduced lunch, and percent of students with disabilities) had limited contributions to explaining the classroom-level variation in the classroom observation ratings. Moreover, some predictors (i.e., percent of minority students, percent of male students, percent of ELLs, and class size) were statistically significant but had very small estimates of impact. Therefore, this dissertation may not provide an implication regarding using these classroom characteristics for the observation rating adjustment in teacher evaluations.

Moreover, a CCREM was used for the analysis in this dissertation because multiple ratings were given by multiple raters per item per classroom, and classrooms were nested within teachers within schools. As discussed earlier, ignoring or misspecifying the cross-classification structure may generate biased fixed effects estimates, standard error estimates, and variance component estimates (Fielding & Goldstein; 2006; Luo & Kwok, 2010; Meyers & Beretvas, 2006; Rasbash & Browne, 2001; Wallace, 2015). Prior research also found that a CCREM is more appropriate to use than other types of advanced models such as a cross-classified multiple membership random effects model (CCMMrem) to control rater bias in estimating teacher effectiveness (Murphy & Beretvas, 2015). Therefore, appropriately accounting for the cross-classification structure is crucial for examining the variation of classroom observation ratings.

Limitations and Suggetions for Future Research

This dissertation highlights issues related to the variation of a teacher's classroom observation ratings across his or her multiple classrooms. However, there are a few meaningful and important issues that are beyond the scope of this dissertation and should be researched in future work. First and foremost, this dissertation did not provide evidence related to whether ignoring classroom-level variation in observation ratings would lead teachers being misclassified in a teacher evaluation system. Whitehurst et al. (2014) found that adjusting the observation scores by controlling for the classroom effects could move some teachers out of their original ranking positions in teacher evaluations. However, Lazarev and Newman (2015) pointed out that the adjustment of observation ratings might not be appropriate if teacher assignment was not random. For example, if less proficient teachers were assigned to lower-performing classrooms or if schools were less successful in retaining effective teachers, then such an adjustment would mask the real comparisons among teachers. Therefore, how to incorporate the observation rating adjustment in teacher evaluations and how this adjustment for contextual factors (e.g., raters, schools, classrooms) would function in rewarding or sanctioning teachers in high-stakes settings could be investigated in future research.

Furthermore, what drives the classroom-level variation in observation ratings is not fully clear. Whitehurst et al. (2014) suggested that it is possible that teachers with challenging students may not perform well or raters tend to assign lower ratings to teachers leading challenging classrooms, regardless of the teachers' actual performance. However, the classroom characteristics (i.e., class size, percent of minority students, percent of male students, percent of ELLs, percent of students eligible for free or reduced lunch, and percent of students with disabilities) had limited contributions to explaining the classroom-level variation in observation ratings. It is possible that there are other sources of contextual influence on the classroom-level variation in observation ratings such as student belief and student knowledge (Bell et al., 2012; Hill et al., 2012). Because observation ratings can be influenced by many factors, what drives the variation of observation ratings across teachers' multiple classrooms needs more investigation in future research.

In addition, this dissertation did not include the student achievement scores as a classroom-level predictor to explain the classroom-level variation of classroom observation ratings. Research indicates that states and districts may adjust classroom observation ratings by controlling for the student achievement level because teachers with better-performing students have unfair advantages to receive higher observation ratings (Whitehurst et al., 2014). However, in the sample of this dissertation, student state test scores were converted to rank-based *z*-scores within district, subject, and grade (White & Rowan, 2014). This dissertation conducted analyses involving all six districts and six grade levels where the student achievement level was not used as a predictor. How student achievement scores can explain the classroom-level variation in observation ratings is a topic that would merit future research.

Another limitation of this dissertation is that the variation of classroom observation ratings may be influenced by the scoring design and observation implementation (Bell et al., 2012; Hill et al., 2012). As an illustration, Hill et al. (2012) found that whether the rater viewed the first 30 minutes of the lesson or the entire lesson could influence the variation of classroom observation ratings. For ratings of the Measures of Effective Teaching (MET) project used in this dissertation, data for each classroom was collected from the first 30 minutes of each lesson, which may not be representative of the entire lesson.

This dissertation also has limitations in terms of its generalizability. First, the MET project collected data from six school districts, and the results of this dissertation may not be generalized to students and teachers from other districts that did not participate in the MET project. It is possible that the classroom-level variation in observation ratings may differ in other districts. Second, this dissertation only analyzed classroom observation ratings based on one instrument, the CLASS. Other instruments, such as the Framework for Teaching (FFT), Mathematical Quality of Instruction (MQI), Protocol for language Arts Teaching Observation (PLATO Prime), may be investigated in future studies regarding the variation of a teacher's classroom observation ratings across multiple classrooms.

REFERENCES

Abbott, M. L., & Fouts, J. T. (2003). *Constructivist teaching and student achievement: The results of a school-level classroom observation study in Washington.* Retrieved from Seattle Pacific University, Washington School Research Center website: http://spu.edu/orgs/research/ObservationStudy-2-13-03.pdf

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher-student interactions in secondary school classrooms: predicting student achievement with the Classroom Assessment Scoring System- Secondary. *School Psychology Review, 42*(1), 76-97.

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Retrieved from http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

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

Beretvas, S. N. (2001). Cross-classified and multiple membership models. In J. J. Hox, & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 313-344). New York, NY: Routledge.

Berliner, D. C. (2014). Exogenous variables and value-added assessment: A fatal flaw. *Teachers College Record, 116*(1). Retrieved from http://www.tcrecord.org/Content.asp?ContentId=17293

Blatchford, P., Bassett, P., & Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools. *Learning and Instruction*, *21*(6), 715-730.

Board, J. (2011). *Classroom Observation-Purposes of Classroom Observation, Limitations of Classroom Observation, New Directions.* Retrieved from http://education.stateuniversity.com/pages/1835/Classroom-Observation.html

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 1*(3), 472-514. doi:10.1214/06-BA117

Bruhwiler, C., & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and Instruction, 21*(1), 95-108. doi:10.1016/j.learninstruc.2009.11.004

Cadima, J., Peixoto, C., & Leal, T. (2014). Observed classroom quality in first grade: Associations with teacher, classroom, and school characteristics. *European Journal of Psychology of Education, 29*(1), 139-158. doi:10.1007/s10212-013-0191-4

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Science, 3*(3), 106-116.

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (No. w17699). Retrieved from http://www.nber.org/papers/w17699

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

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1), 1-44. Retrieved from http://epaa.asu.edu/ojs/article/view/392/515

Darling-Hammond, L. (2010). *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching?* Washington, DC: Center for American Progress. Retrieved from http://files.eric.ed.gov/fulltext/ED535859.pdf

Darling-Hammond, L. (2012). *Creating a comprehensive system for evaluating and supporting effective teaching.* Standford, CA: Standford Center for Opportunity Policy in Education.

Dreeben, R., & Barr, R. (1988). Classroom composition and the design of instruction. *Sociology of Education, 61*(3), 129-142. doi: 10.2307/2112622

Evertson, C. M., Anderson, C. W., Anderson, L. M., & Brophy, J. E. (1980). Relationships between classroom behaviors and student outcomes in junior high mathematics and English classes. *American Educational Research Journal, 17*(1), 43-60. doi: 10.3102/00028312017001043

Fennema, E., & Peterson, P. L. (1985). Autonomous learning behavior: A possible explanation of sex-related differences in mathematics. *Educational Studies in Mathematics, 16*(3), 309-311. doi:10.1007/BF00776738

Fielding, A., & Goldstein, H. (2006). *Cross-classified and multiple membership structures in multilevel models: An introduction and review* (Research Report No. 791). Retrieved from http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/cross-classified-review.pdf

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock, & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269-314). Greenwich, Connecticut: Information Age Publishing.

Garrett, R., & Steinberg, M. P. (2015). Examining Teacher Effectiveness Using Classroom Observation Scores Evidence From the Randomization of Teachers to Students. *Educational Evaluation and Policy Analysis, 37*(2), 224-242. doi:10.3102/0162373714537551

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://files.eric.ed.gov/fulltext/ED521228.pdf

Goldstein, H. (2003). *Multilevel statistical models*. (3rd ed.). New York, NY: Hodder Arnold.

Hamre, B.K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4000 U.S. early childhood and elementary classrooms*. Retrieved from http://fcd-us.org/sites/default/files/BuildingAScienceOfClassroomsPiantaHamre.pdf

Hansen, M., Lemke, M., & Sorensen, N. (2013). *Combining multiple performance measures: Do common approaches undermine districts' personnel evaluation systems?* Washington DC: American Institutes for Research. Retrieved from http://www.air.org/sites/default/files/downloads/report/Combining_Multiple_Performance_Measures_0.pdf

Hattie, J. A. (2002). Classroom composition and peer effects. *International Journal of Educational Research, 37*(5), 449-481.

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M.,…Lynchm K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment, 17*(2-3), 88-106. doi: 10.1080/10627197.2012.715019

Hill, P. W., & Goldstein, H. (1998). Multilevel modeling of educational data with cross-classification and missing identification of units. *Journal of Educational and Behavioral statistics,23*(2), 117-128. doi: 10.3102/10769986023002117

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Hull, J. (2013). *Trends in teacher evaluation: How states are measuring teacher performance.* Center for Public Education. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Trends-in-Teacher-Evaluation-At-A-Glance/Trends-in-Teacher-Evaluation-Full-Report-PDF.pdf

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

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper. Pdf

Kelcey, B., McGinn, D., & Hill, H. (2013). Measurement of classroom teaching quality with Item Response Theory. Paper presented at the Society for Research on Educational

Effectiveness, Washington, DC.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher, 39*(8), 591–598. doi: 10.3102/0013189X10390804

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746-759. doi: 10.1177/0013164496056005002

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61*(2), 213-218. doi: 10.1177/00131640121971185

La Paro, K. M., Hamre, B. K., Locasale-Crouch, J., Pianta, R., Bryant, D., Early, D., ... Burchinal, M. (2009). Quality in kindergarten classrooms: Observational evidence for the need to increase children's learning opportunities in early education classrooms. *Early Education & Development, 20*(4), 657–692.

Lazarev, V., & Newman, D. (2015). How Teacher Evaluation is affected by Class Characteristics: Are Observations Biased?. Social Science Research Network. Retrieved from http://dx.doi.org/10.2139/ssrn.2574897

Leo, S. F., & Lachlan-Haché, L. (2012). *Creating summative educator effectiveness scores: Approaches to combining measures.* Washington, DC: American Institutes for Research. Retrieved from http://educatortalent.org/inc/docs/Creating%20Summative%20EE%20Scores_FINAL.PDF

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics, 27*(3), 255-270. doi: 10.3102/10769986027003255

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*(4), 514-534.

Luo, W., & Kwok, O. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral Research*, *44*(2), 182-212.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., 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.

Marzano, R. J., & Toth, M. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement*. Alexandria, Virginia: Association for Supervision and Curriculum Development.

McCoach, D. B., & Black, A. C. (2008). Evaluation of model fit and adequacy. In A. A. O'Connell, & D. B. McCoach (Eds.), *Multilevel modeling of educational data*. (pp.245-272). Greenwich, Connecticut: Information Age Publishing.

McGuinn, P. (2012). *The state of teacher evaluation reform: State education agency capacity and the implementation of new teacher-evaluation systems.* Washington, DC: Center for American Progress. Retrieved from https://cdn.americanprogress.org/wp-content/uploads/2012/11/McGuinn_TheStateofEvaluation-1.pdf

Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research*, *41*(4), 473-497.

Mihaly, K., & McCaffrey, D. F. (2014). Grade-level variation in observational measures of teacher effectiveness. 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. 9-49). San Francisco, CA: Jossey-Bass.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching.* Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

Morgan, G. B., Hodge, K. J., Trepinski, T. M., Anderson, L. W. (2014). The stability of teacher performance and effectiveness: Implications for policies concerning teacher evaluation. *Education Policy Analysis Archives, 22*(95), 1-21. Retrieved from http://files.eric.ed.gov/fulltext/EJ1050120.pdf

Murphy, D. L., & Beretvas, S. N. (2015). A comparison of teacher effectiveness measures calculated using three multilevel models for raters effects. *Applied Measurement in Education, 28*(3), 219-236. doi:10.1080/08957347.2015.1042158

National Council on Teacher Quality. (2015). *State of the states 2015: Evaluating teaching, leading and learning.* Washington DC. Retrieved from http://www.nctq.org/dmsView/StateofStates2015

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15*(5), 625-632. doi:10.1007/s10459-010-9222-y

O'Connell, A. A., & Reed, S. J. (2012). Hierarchical data structures, institutional research, and multilevel modeling. *New Directions for Institutional Research, 2012*(154), 5-22.

Pakarinen, E., Lerkkanen, M. K., Poikkeus, A. M., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., & Nurmi, J. E. (2010). A validation of the classroom assessment scoring system in Finnish kindergartens. *Early Education and Development, 21*(1), 95-124.

Park, Y. S., Chen, J., & Holtzman, S. L. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. 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. 381-414). San Francisco, CA: Jossey-Bass.

Partee, G. L. (2012). *Using multiple evaluation measures to improve teacher effectiveness: State Strategies from round 2 of No Child Left Behind Act waivers*. Washington, DC: Center for American Progress. Retrieved from https://www.americanprogress.org/wp-content/uploads/2012/12/MultipleMeasures-2-INTRO.pdf

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. doi:10.3102/0013189X09332374

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System: Manual k-3 version*. Baltimore: Paul Brookes Publishing.

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

R Development Core Team. (2014). R: A language and environment for statistical computing (Version 3.1.2) [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rasbash, J., & Browne, W. J. (2001). Modeling non-hierarchical structures. In A. H. Leyland and H. Goldstein (Eds.), *Multilevel modeling of health statistics* (pp. 93-105). Chichester, UK: Institute of Education.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral*

Statistics, 19(4), 337-350. doi: 10.3102/10769986019004337

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

Reyhner, J. (1991). The challenge of teaching minority students: An American Indian example. *Teaching Education*, *4*(1), 103-112.

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

Rjosk, C., Richter, D., Hochweber, J., Lüdtke, O., & Stanat, P. (2015). Classroom composition and language minority students' motivation in language lessons. *Journal of Educational Psychology*, *107*(4), 1171-1185.

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252.

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 57-67. doi:10.1023/A:1007999204543

Sandilos, L. E., DiPerna, J. C., & Family Life Project Key Investigators. (2014). Measuring Quality in Kindergarten Classrooms: Structural Analysis of the Classroom Assessment Scoring System (CLASS K–3). *Early Education and Development*, *25*(6), 894-914.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the classroom observations of student-teacher interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly, 27*(2), 316-328. doi:10.1016/j.ecresq.2011.09.004

Snijders, T. B., & Bosker, R. R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, *24*(3), 97–117. doi: 10.1257/jep.24.3.97

Steele, J. L., Hamilton, L. S., & Stecher, B. (2010). *Incorporating student performance measures into teacher evaluation systems.* RAND Corporation. Retrieved from http://www.rand.org/content/dam/rand/pubs/technical_reports/2010/RAND_TR917.pdf

Steenbergen, M. R., & Jones, B. S. (2002). Modeling multilevel data structures. *American Journal of Political Science, 46*(1), 218-237. doi: 10.2307/3088424

Stuhlman, M. W., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2010). *A practitioner's guide to conducting classroom observations: What the research tells us about choosing and using observational systems.* The University of Virginia Center for Advanced Study of Teaching and Learning. Retrieved from http://curry.virginia.edu/uploads/resourceLibrary/CASTL_practioner_Part2_single.pdf

Theall, K. P., Scribner, R., Broyles, S., Yu, Q., Chotalia, J., Simonsen, N., ... Carlin, B. P. (2011). Impact of small group size on neighbourhood influences in multilevel models. *Journal of epidemiology and community health*, *65*(8), 688-695. doi:10.1136/jech.2009.097956

The National Center on Quality Teaching and Learning. (2012). *Understanding and using CLASS for program improvement.* Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/tta-system/teaching/docs/class-brief.pdf

The National Center on Quality Teaching and Learning. (2013). *Improving teacher-child interactions: Using the CLASS in head start preschool programs.* Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/tta-system/teaching/docs/using-the-class.pdf

Trueba, H. T. (1988). Culturally based explanations of minority students' academic achievement. *Anthropology & Education Quarterly, 19*(3), 270-287. doi: 10.1525/aeq.1988.19.3.05x1565e

U.S. Department of Education (2015a). *The RESPECT Project Vision Statement.* Retrieved from http://www.ed.gov/teaching/national-conversation/vision

U.S. Department of Education (2015b). *ESEA Flexibility*. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

U.S. Department of Education (2015c). *Every Student Succeeds Act (ESSA).* Retrieved from http://www.ed.gov/essa?src=policy

Wallace, M. L. (2015). *Modeling cross-classified data with and without the crossed factors' random effects' interaction* (Doctoral dissertation). Retrieved from https://repositories.lib.utexas.edu/handle/2152/31010

Waxman, H. C., & Huang, S. L. (1999). Classroom observation research and the improvement of teaching. In H. C. Waxman & H. J. Walberg (Eds.), *New directions for teaching practice and research* (pp.107-129). Berkeley, CA: McCutchan.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: concepts, issues, and applications* (pp. 56-75). Thousand Oaks, CA: Sage.

White, M., & Rowan, B. (2013). *User guide to measures of effective teaching longitudinal database (MET LDB).* Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Whitehust, G. J., Chingos, M. M, & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: Brown Center on Education Policy at Brookings. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2014/05/13-teacher-evaluation/evaluating-teachers-with-classroom-observations.pdf

Yuan, K., McCaffrey, D. F., & Savitsky, T. D. (2013). *Analyzing the Factorial Structure of the Classroom Assessment Scoring System-Secondary Using a Bayesian Hierarchical Multivariate Ordinal Model*. Society for Research on Educational Effectiveness.