THE INFLUENCE OF CONTENT AND OTHER FACTORS ON MEASURES OF TEACHER QUALITY: EVIDENCE FROM TEACHERS' ENGLISH LANGUAGE ARTS AND MATHEMATICS INSTRUCTION

By

Sihua Hu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Mathematics Education—Doctor of Philosophy

2016

ABSTRACT

THE INFLUENCE OF CONTENT AND OTHER FACTORS ON MEASURES OF TEACHER QUALITY: EVIDENCE FROM TEACHERS' ENGLISH LANGUAGE ARTS AND MATHEMATICS INSTRUCTION

By

Sihua Hu

This study uses data from the Measures of Effective Teaching (MET) project to examine the differences in teachers' observational measures across content under various contexts: 1) the principled choice of the instrument, 2) the score aggregation methods to generate the final ratings, and 3) the performance frameworks under which teachers are categorized. Specifically, this study examines whether the same teachers' observational measures in distinctive subjects (ELA vs. mathematics) as well as subject areas within mathematics (e.g., Algebra vs. Geometry) are different, and hence influencing their evaluation results non-trivially. For the generalist teachers, this study finds that there are statistical differences as well as practical differences in the same teachers' observational measures between ELA and mathematics. Such differences are present for both the generic instruments and the subject-specific instruments, and do not depend on the grade level.

For the mathematics teachers, this study compares the consistency of their observational measures between the two generic instruments, and between the generic instruments and a math-specific one. The results show that the two generic instruments have much higher consistency with each other than with the subject-specific one respectively. Moreover, almost none of the differences between the same mathematics teachers' observational measures across unlike subject areas are statistically significant. Under the relative performance framework, however,

analyses using the rank scores demonstrate a large volatility between teachers' two observational measures across areas of mathematics.

In conclusion, there is a lack of consistency between the same teachers' observational measures when they are observed in diverse content, especially different subjects, across all three contexts: instrument choice, score aggregation methods, and performance framework for categorization. The consistency, or the lack thereof, when teachers are observed in different areas of mathematics, however, have some associations with the performance framework to put teachers into quality categories. These findings all together suggest the need to take into consideration the content of the lessons the evaluator chooses to observe, and provide some empirical evidence on the implementation of the observation component in current teacher evaluation systems.

ACKNOWLEDGEMENTS

I want to express my greatest appreciation to my committee co-chairs—Dr. Robert Floden and Dr. Vincent Melfi, as well as my other two committee members—Dr. Kristen Bieda and Dr. Sharon Senk. Without their support and insights, I would not be able to complete this dissertation.

I am grateful that I joined the Program in Mathematics Education at Michigan State University six years ago. Throughout my doctoral study, I have grown professionally as well as emotionally with the support from our local mathematics education community. I thank all the faculty members that I have taken a course and/or engaged in research with in the College of Education and the College of Natural Sciences. Particularly, I want to thank my first year advisor Jennifer Kaplan to spark my interest in statistics education. Without her encouragement, I would not pursue my master degree in statistics along the way. I want to give my appreciation to my all-time advisor and dissertation co-chair—Dr. Vincent Melfi, who have guided me to navigate through my doctoral study. Utmost gratitude goes to my research supervisor, practicum and dissertation co-chair—Dr. Robert Floden, without whom I would not have found my research interest in classroom research. Most importantly, he has taught me how to think critically as an educational researcher by being a role model. I also want to extend my gratitude to Dr. Robert Floden's collaborator—Dr. Alan Schoenfeld at University of California-Berkeley. He has been an inspiring mentor to work with during my four years on the Algebra Teaching Study project. I am indebted to Dr. Kristen Bieda who invited me to join the Study of Elementary Mathematics Instruction project. She has provided me the opportunities to collaborate with people from other fields in education, and to start practicing leadership within a large research team.

I am thankful to have our program coordinator Lisa Keller, who takes care of everything for our graduate students. I am also fortunate to have a supportive group of colleagues to inspire me and motivate me along the way. Many thanks to my writing group members who have help me stay focused and organized in my writing, not limited to my dissertation.

Last but not least, I am deepest grateful to my boyfriend, Xun Wang, for his love and support during this process. The end of a journey is the beginning of another one, and I am ready to embark on new adventure with all the previous things that I have learned at MSU.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. Purpose of The Study: The Role of Content in Teaching and Its Manifestation in Teachers' Observational Scores

For a long time, scholars viewed teaching as a generic activity that transcended the content (e.g., Gage, 1978). Shulman (1986, 1987), who was among the earliest educators to note the importance of subject matter in teaching, argued the necessity of attending to the different types of teacher knowledge beyond simply pedagogical. Educators who followed Shulman's line of logic argued that teaching is a subject-specific activity; mathematics teachers possess content knowledge and pedagogical content knowledge that apply exclusively to the teaching of mathematics (Ball, Thames, & Phelps, 2008; Hill, Schilling, & Ball, 2004), which are different from knowing advanced mathematical content and general pedagogical skills. Subject matter thus serves as a pivotal context around which teachers organize and facilitate different curricular and classroom activities (Stodolsky, 1988; Stodolsky & Grossman, 1995).

Teaching is a complex activity that involves more than interactions with students when delivering a lesson. The educational research community has been examining teaching in the classroom for decades. Despite the long history, observational protocols were mostly used by scholars for the purpose of classroom research, such as to understand teachers' practices in relation to student outcomes under the process-product paradigm. Not until the last decade, however, were observational protocols utilized for a wider range of purposes, not limited to research and professional development, but also for teacher evaluation under the current policy climate.

Many districts nowadays are using some types of observational protocols to measure teachers' instructional practices in the classroom in order to fulfill their accountability

responsibility. Teachers are facing more pressure on top of their teaching activities. They are being observed and evaluated by experts or administrators in their own classrooms, and are held accountable for their students' achievement in standardized tests. Teaching is hard, and evaluation should not make teachers' lives harder by introducing unfair judgments and unconscious biases as the inferences that stakeholders made based on the evaluation results is directly related to the interests of teachers. Accordingly, it is essential to establish a validation program regarding the use of observational protocols to capture the construct of teacher quality within states' accountability framework. In the end, a valid and reliable measure of teacher quality is not just about teachers, but it is also directly tied with the quality of schools, districts, and the whole education enterprise in the US.

This study takes advantage of the large data set and score-ready lessons in the Measures of Effective Teaching (MET) project to study teacher quality as captured by various observational protocols. The MET project is a research project funded by the Gates Foundation that actively seeks empirical evidence to link various aspects of teacher quality to student achievement gains. The project collaborated with six districts and more than 3,000 teachers to collect data on classroom observations, student test scores, background information, and surveys from relevant actors in the education system. The MET project differs from most existing studies of teachers not just because of its large number of participants, but also because of the various measures used on the same population of teachers, including different types of classroom observational protocols to characterize the very same population.

In particular, this study explores the role played by content in teaching activities and their manifestation in the observational scores obtained from different protocols. Content refers to not only the subject, such as mathematics versus English Literature Arts (ELA), as some previous

studies meant, but also the subject areas within a discipline, which is mathematics in this study. In the mathematics community of higher education (maybe apart from those intersecting areas such as Algebraic Geometry), people would say without hesitation that the teaching practices and goals of a pure algebraist such as J.J. Sylvester are quite different from, if not completely opposite of, the teaching practices and goals of a geometrician such as Felix Kline (Parshall, 2003). Mathematicians of different areas are dealing with distinct mathematical objects on a daily basis, and so do mathematics teachers when they teach different subject areas. Accordingly, my assumption is that the mathematics teachers' teaching practices in different areas are qualitatively different, even when taught by the same teacher. This study uses the MET data to examine whether such qualitative differences in knowledge and practices also manifest themselves quantitatively in observational measures.

Additionally, this study can also be seen as a study regarding the reliability and validity issues on the use of various observational protocols that are widely used in the K-12 classrooms. As argued by Kane (2001, 2012), validity is never the validation of the instrument itself; rather, it is the proposed interpretations of the scores of the measurement that is to be validated. That is to say, if observational protocols are to be used for teacher evaluation, the users of these protocols have the obligations to validate their interpretations of the results from using the protocols, and consider the consequences of the interpretations with respect to decision procedures that affect the teachers. The purpose of this study is to examine whether observational scores from different protocols are sensitive to subject (ELA vs. mathematics) and subject areas within mathematics under relevant contexts of teacher evaluation, and to discuss the consequences of such sensitivity in the policy climate. If various observational protocols identify the same teacher's quality differently because of the content observed and/or because of

other contextual factors such as the performance framework used to categorize teachers, then what are the political considerations of ignoring such differences in teacher evaluation? This study examines those potentially influential factors in teachers' observational scores and evaluation results to answer this general question.

1.2. Research Questions

Using the MET data, this study asks three specific research questions as follows:

1. For the generalist teachers in Grades 4-6 who teach both ELA and mathematics, to what extent are their observational scores different, as measured by various protocols in the MET data?
    - To what extent are their teacher quality measures different across subjects, as assessed by the same generic observational instruments? Generic instruments examined include Framework for Teaching (FFT) and Classroom Assessment Scoring System (CLASS).
    - To what extent are their teacher quality measures different across subjects, as measured by subject-specific observational instruments respectively? Subject-specific instruments are Protocol for Language Arts Observation (PLATO) for ELA, and Mathematical Quality of Instruction (MQI)) for mathematics.
2. For the mathematics teachers in Grades 4-9, to what extent are their teacher quality measures different, as assessed by various observational instruments in the MET data?
    - To what extent do the two generic instruments measure teacher quality in mathematics differently from each other?
    - To what extent do the subject-specific and one of the generic instruments measure teacher quality in mathematics differently from each other?

3. For the mathematics teachers in Grades 4-9, to what extent do their teacher quality measures differ across subject areas within mathematics, as assessed by the generic and math-specific observational instruments? Subject areas in mathematics are defined by the domain specified in the Common Core State Standards (CCSS, 2010), including: Numbers & Operations; Algebra & Algebraic Thinking (including expressions, equations, functions, and high school algebra); Geometry; and Statistics & Probability.

1.3. Significance of the Study

This large scale study of teacher quality examining multiple measures of teaching practices and relevant contexts allows for a better understanding and a broader consideration of the influential factors in teacher evaluation. The results from this study can be used by educational researchers, administrators, and policymakers to inform about the implementations of teacher evaluation systems across the nation. If teacher quality is indeed sensitive to the disciplines and/or to areas within the discipline, such variability should be brought to the conscious level of teachers and educational researchers so that they can work together to create a common professional knowledge base of high leverage practices. Also, the evaluators of teachers should take into consideration those potentially influential factors when s/he observes teachers teaching particular lessons. If teachers' observational ratings are not consistent across instruments and ways of using the instruments and the scores, teacher evaluation systems should be aware of the potential bias in inferences resulted from these inconsistency, and make evidence-based decisions on the implementations of an observational system to evaluate teachers.

In sum, this study contributes to body of literature on measuring and understanding teacher quality in light of teacher evaluation and the policy discourse within and around it. By investigating characteristics of teacher quality measures in different contexts, this study hopes to provide a better understanding of observation protocols to be used in the classrooms, especially for the purpose of teacher evaluation.

CHAPTER 2

BACKGROUND AND CONCEPTUAL FRAMEWORK

In order to situate this study in the larger body of educational research and illustrate the relevance of the results for policy, it is important to know the literature and political practices around the teacher quality construct. In this section, I first outline the multiple dimensions of this construct as conceptualized in prior literature. Then I introduce the current practices in most teacher evaluation systems and identify assumptions of practices that are unexamined. Next, I turn my focus to the research-based assessments of teacher quality—observational protocols. The assessments reviewed focus on the ones used in the MET study as means to make inferences about teacher quality from their teaching performance for teacher evaluation. I describe some of the validity and reliability studies in existence for these assessments, and point out the gaps in their research programs for validation. Using a framework of construct validation, I provide supports for the importance of filling such gaps in the validation processes. I outline my hypotheses on the potentially influential factors in these classroom assessments of teachers, and discuss past research that built up my hypotheses. Lastly, I present a conceptual framework to summarize the relationships among components of teacher quality and teacher evaluation systems in order to situate the contributions of this study.

2.1. Literature Review on Teacher Quality

Before delving into the measures of teacher quality, it is essential to unfold how researchers have conceptualized theoretically this underlying construct of teacher. Teacher quality is a broadly defined construct relating to teachers and their professional activities, but it is not directly observable, nor is it a static trait of teachers. Accordingly, there is no universal consensus on the characteristics of quality teachers and quality teaching.

Many researchers, such as Wenglinsky (2000) and Kennedy (2004), argued that teacher quality is a multi-faceted construct which encompasses many aspects. Wenglinsky (2000) summarized three types of teacher quality measures: teacher inputs (e.g., years of experience and education level), classroom practices, and professional development. He contended that previous research and policy have primarily focused on the first type of measures, which are the non-classroom aspects of teacher quality. The classroom aspects of teacher quality, especially teachers' classroom practices, however, are a stronger predictor of student improvement in terms of their learning outcomes.

Kennedy (2004) provided a more comprehensive and detailed framework of teacher quality by incorporating teacher effectiveness as well as teachers' affective and motivational factors. She summarized four main aspects of teacher quality that have been examined by the research community: 1) qualifications, 2) effectiveness, 3) quality of practices, and 4) orientation. Most research on teacher quality can be seen as the examination of some combinations of the above-mentioned aspects of teacher quality. However, she noted that these aspects have not distinguished themselves from one another and have been used interchangeably in the literature. For example, some researchers directly defined teacher quality in terms of student achievement (e.g., Rivkin, Hanushek, & Kain, 2005) and related student outcomes with other teacher experience variables such as first year teaching and courses taken in teacher preparation. In contrast, other researchers defined teacher quality in terms of teacher qualifications and related this defined teacher quality to student achievement (e.g., Darling-Hammond, 2000; Rice, 2003). In both cases, even though teacher quality was defined differently, the researchers were examining the relationship between their choices of teacher qualifications and teacher effectiveness, which are two aspects of the multi-faceted teacher quality.

Another taxonomy of teacher quality put forward by Kennedy (2008) consists of three categories: 1) personal resources, 2) performance, and 3) effectiveness. This taxonomy is in essence similar to the four aspects of teacher equality described above, with the combination of qualifications with orientation into personal resources that teachers bringing to the profession. Moreover, details of each sub-dimension are added in this later version, including pertaining behaviors and traits, and examples of the different assessments to measure them (see, Table 1).

Table 1: Taxonomy of teacher quality and its sub-dimensions (Kennedy, 2004; 2008)

Aspects of Teacher Quality		**Subdivisions**
Qualifications	Personal Resources	Beliefs, attitudes, and values
		Personal traits
Orientation		Knowledge, skills, and expertise
		Credentials
Quality of Teaching	Performance	Practices within the classroom
		Lesson planning
		Collaborating with colleagues
		Non-academic support for the students
Teacher Effectiveness	Effectiveness	Raising student scores on standardized achievement tests
		Raising student scores on cognitive demanding assessments
		Motivating students
		Fostering students' sense of responsibility and social concern

According to Kennedy, these lists of behaviors and traits are not meant to be exhausted, because each dimension is a sub-construct that can mean many different things according to how people conceptualize it. Take effectiveness as an example, the most common but narrow definition of effectiveness is students' scores in achievement tests, which include standardized tests and tests that aim to assess higher thinking order and problem solving. Student achievement

in these tests is a proxy of student learning outcomes depending on the content and skills covered by the assessments. At the same time, student learning outcomes also include their orientations such as beliefs and attitudes resulting from schooling. In some contexts, teacher effectiveness may even be conceptualized in terms of other non-student outcomes. For example, in teacher induction and mentoring research, teacher effectiveness is used to describe the resulting culture of the school and teachers' local community from high quality induction program and mentoring (Strong, 2008). In other words, the effectiveness of a teacher can be defined as his or her influence on the local community and other (new) teachers.

The importance of being explicit and precise in our uses of the term teacher quality is not apparent. We have an idea that the different aspects of teacher quality delineated above are interconnected, but even this assumption should not go unexamined. The more important questions are to what extent are they related to one another and in what ways are they related so that changes in one aspect lead to changes in another. Moreover, as claimed by Kennedy (2004), both researchers and policymakers need to know what they are referring to as teacher quality in order to "improve our ability to measure it [teacher quality], improve it, or reward it (p. 60)."

2.2. Recent Teacher Evaluation Practices in the Nation

This part of the literature review uses an illustrative example of a teacher evaluation system to introduce some common practices of the observation component. By highlighting the procedures in the enactment of the observation component, I identify three understudied factors not emphasized in the current practices: 1) the content of the lessons being observed, 2) the principled choice of the instrument, and 3) the method of generating composite scores to represent teacher quality. I argue that to examine the influence of these three factors in this study,

it helps inform the teacher evaluation practices and provide rationale to support certain ways of implementing the observational protocols.

Since 2009, the design and implementation of teacher evaluation systems have been on every state's policy agenda in order to qualify for the now defunded Race to the Top grant and No Child Left Behind waivers under the Obama administration. According to the National Council on Teacher Quality (Doherty & Jacobs, 2015), by 2015, most states, except for five[1], have incorporated teacher evaluation in their state policy, and about half of them have used the evaluation results for tenure or dismissal decisions. Classroom observation was a component in every state's evaluation system at the time of 2013, as reported by Center for Public Education (Hull, 2013). As teacher evaluation reform is highly volatile to changes in many states, nine of the states[2] no longer specify observation in their teacher evaluation policy by the end of 2015 (Doherty & Jacobs, 2015). Still, the majority of the states have specified the use of observational measures to evaluate teachers, and for those states where observations were not mandated, it was still common for districts to incorporate the observation component.

Although observation of teachers has been a less debatable component than measures of student outcomes (e.g., Value-added Model Scores and Student Growth Percentiles) in teacher evaluation systems, there is few specifications on the practices and processes of using observation. In this section, I illustrate a teacher evaluation system—IMPACT—adopted by the District of Columbia (D.C.) as an example of implementing classroom observations to evaluate teachers. I chose the example of D.C. as it is one of the pioneers in educational reforms for the last decade, and has been experimenting with teacher evaluation practices ahead of many other

---

[1] California, Iowa, Montana, Nebraska and Vermont.
[2] California, District of Columbia, Kansas, Montana, New Hampshire, North Dakota, Texas, Vermont, and Wyoming.

states. Hence, the IMPACT system has been closely watched and even imitated by most parties in the education community. In particular, it has been the subject of many research studies and reports (e.g., Dee & Wyckoff, 2015; Headden, 2011) as an example to provide evidence on what works or does not work in teacher evaluation systems. IMPACT underwent many changes over the recent years in response to the shift of policy discourse and many other external and internal factors. Herein I focus on the system implemented in the public schools in D.C. for the 2015-2016 school year, which is the most recent version of the enacted teacher evaluation for the area.

As a performance and incentive based evaluation system, IMPACT differentiates the number of observations based on teachers' stages in their career. According to the district's guidebook, the system places teachers in five developing stages. Teachers at the earlier two stages receive four formal observations and one informal observation yearly. Among these observations, administrators are responsible for two formal observations and the informal one, while master educators are responsible to conduct the other two formal observations. The calculation algorithm employed by IMPACT is that the observational scores are averaged in each dimension in order to get the final composite score for each lesson, and the extreme aggregated score (one point difference on a 4-point scale in comparison to other evaluators' scores) is dropped.

From the description, one can see that the observation component in IMPACT focuses on the frequency of observations and the backgrounds of evaluators, which coincide with most of the research efforts in this area within the research community. Past studies on the use of observational measures in teacher evaluation consistently recommend multiple observations, and multiple evaluators to make the scores more reliable (Hill, Kapitula, & Umland, 2011; Ho &

Kane, 2013). The bias minimized in these two practices is mostly the sample insufficiency and rater bias among many other influential factors, which directly addresses the reliability issues.

Also, just like many other states, D.C.'s teacher evaluation system uses generic and subject-free rubrics for observation, and there is no specification on the content of the observations that the evaluators should choose to observe. Some states develop their own observational rubrics internally and trust the scores in implementation. Other states, such as Michigan (Michigan Council for Educator Effectiveness, 2013), chose research-based observational instruments and made recommendations to the districts to let them choose from a small set of protocols. In both situations, the observational protocols do not have a foothold in subject-specific practices explicitly. Interestingly, there is a trend in the research community to develop and utilize subject-specific observational protocols to measure instructional quality (Schlesinger & Jentsch, 2016), while districts and states uniformly used generic and content-free rubrics to evaluate teachers. The gap between the preference of the research community and the states/districts are not yet addressed by current studies, and neither preference is supported by empirical evidence.

Lastly, another prevailing practice adopted by IMPACT and many other states' teacher evaluation systems is to generate a univariate composite score by averaging across dimensions of the rubrics to represent teacher quality. There are several assumptions that go unexamined with this approach. The main underlying assumption is that the construct of teacher quality can be broken down into multiple uncorrelated parts, and each part has equal weight in accounting for the construct. Whether such assumption holds and whether the composite scores of averaging all dimensions are are valid and reliable measures of teacher quality depends on the instrument itself, and possibly many other contexts. There needs to be empirical evidence within the data

collected from the teacher population on whom the evaluation is performed in order to examine some of the assumptions. Recognizing the potential problems of generating composite scores with regard to particular instruments, researchers advocate the use of factor analysis. They claim that this methodology can uncover the systematic relationship among dimensions of the observational rubrics in order to justify one's use of the scores to make meaningful interpretations (e.g., Garrett & Steinberg, 2014; Kane, Taylor, Tyler, & Wooten, 2011). Future chapters elaborate on the two methods and examine the consequences of using them with the data in this study.

In summary, the current practices of conducting observations do not emphasize the content of the lessons that teacher teaches as the evaluation happens, the choice of the tool used to observe teachers, and the method to get the observational ratings as indicators of teacher quality. For these three factors largely neglected, the first one is an arbitrary choice of the evaluator, but the second and third are fixed by the state. Some investigations into these three factors are needed to support and to improve current practices of using observations to evaluate teachers. In the next section, I turn to various assessments to measure different aspects of teacher quality, with a focus on the assessments to measure classroom processes that are included in the MET study.

2.3. Assessments of Teacher Quality

This section first introduces the general background of various assessments of teacher quality, and then transitions to the four observational protocols that are the focus of this study. Past literature addressing the validity and reliability issues of these observational protocols are also reviewed to identify what has been done and what is left to do.

There is a plethora of assessments to measure the various aspects of teacher quality, considering it a multi-faceted construct. The form of assessments includes but not limited to: paper and pencil tests, questionnaires and surveys, interviews, portfolios, self-reported data, and classroom observations. All these assessments are used at different time of a teacher's professional life for specific purposes even though they are all meant to capture some information of the underlying quality that the teacher possesses. Teacher licensure examination, courses taken in teacher preparation, degree, and major are used to assess the qualifications of a teacher candidate; interviews are used to assess a job candidate's personality and his or her fit to the district; classroom observation protocols are used to assess teaching in the context of professional development or annual evaluation. Assessments come from various theoretical frameworks, and are developed for different purposes, especially for the case of the classroom assessments—observational protocols.

2.3.1. Classroom Assessments of Teacher Performance

There is a long history of research using classroom observations, but the protocols[3] or tools used traditionally have changed enormously. Strong (2008) distinguished classroom observational protocols by the amount of inferences that an observer has to make. He contended that there is low inference measure, which is a checklist of prescribed teacher behaviors that only asks an observer to record the counts of each item on the list; in addition, there is high inference

[3] Many researchers distinguish between observational protocol (or observational system) from observational instrument (Boston, Bostic, Lesseig, & Sherman, 2015; Hill, Charalambous, & Kraft, 2012). They consider the observational instrument as a part of the larger observational protocol/system that includes the whole package on using the tool and generating the final ratings. In this paper, I use observational protocol to refer to both the instrument itself and the methods to get aggregated scores within the MET data, and use observational instrument to refer to the tool by itself. The whole observational protocol/system pertains to a particular type of assessment of teacher quality.

measure, which is a coherent rating system that requires an observer to make inferences from a series of classroom events. Most of the classroom observational protocols nowadays pertain to the high inference measure category. They are also both summative and formative in nature so that the same tool can be used for multiple purposes. For teacher evaluation, the classroom protocols are used as summative assessments to get an evaluation of teaching performance. For professional development, the use of classroom protocols is mainly for formative assessment and to provide feedback to improve teaching. For research, both functions have been used widely.

Generally speaking, classroom observational protocols can be divided into two categories: the generic and the subject-specific protocols. Generic protocols tend to focus on the general classroom environment, including classroom culture and norms, as well as management. Despite their rubrics on instruction, the focus is not subject specific (e.g., English Language Arts vs. mathematics) practices, let alone subject-area specific (e.g., Algebra vs. Geometry in mathematics) practices. Thus, they can be used in classrooms across a variety of content, and they also may have some variations in their rubric versions for different grade levels. In contrast, subject-specific protocols are used in one particular subject, or in two closely related subjects, like science and mathematics. The rubrics generally incorporate some subject-specific expectations and specialized instructional practices for ratings. For example, in many observational protocols for mathematics, usually there are some dimensions/rubrics related to teachers and/or students' explanations of their mathematical thinking. Under the rationale that subject-specific protocols conceptualize and measure subject-specific instruction with specialized knowledge in the field, there is a trend in recent years to advocate more uses of these protocols in the classroom research (Schlesinger & Jentsch, 2016). These researchers argued that these subject-specific protocols can direct the attention to classroom processes that are distinct

from general pedagogy and management. These perspectives from which the argument is rooted, however, are research and professional development oriented rather than out of the practical considerations for teachers and administrators.

The MET project only directly focused on the performance/quality of practices aspect of teacher quality. Particularly, the protocols used in the study were subsetted and adapted to some extent to reliably rate teaching practices and teacher-student interactions in the classroom, rather than to capture the planning and preparation (pre-active) and the reflection and refinement of practices (post-active) domains of teaching activities (Strong, 2008). The four observational protocols used in the MET study are high inference assessments that are developed from certain educational research paradigms. The purpose and context for and in which they are proposed to be used, and the theoretical frameworks from which they are built on, are not all the same. The protocols are: Framework for Teaching (FFT), Classroom Assessment Scoring System (CLASS), Protocol for Language Art Observation (PLATO), and Mathematical Quality of Instruction (MQI). The first two protocols are generic observational protocols that were used to score any lesson in the MET study, while the latter two tools are subject-specific protocols that were used in either English Language Art (ELA) or in mathematics lessons. In the following section, I describe the rubrics of each instrument to present what each of them is trying to capture in the classroom processes. How they were operationalized in the MET study is detailed in Chapter 3.

2.3.1.1. Framework for Teaching (FFT)

The FFT protocol was developed and used as a multi-purpose tool, and it has been widely used as a professional development and teacher evaluation tool across the states. As claimed by the developers, FFT is grounded in the constructivist view of learning (Danielson, 2007). Accordingly, the aspects of teacher quality that it measures focus on instruction that would lead

to student-centered constructivist learning. The protocol originally includes ratings in the area of pre-lesson planning and general professional responsibilities. The FFT instrument used in the MET study, however, only includes the domains of Classroom Environment and Instruction, with four dimensions in each domain (see Table 2) and detailed rubrics that describe evidence for the dimension at each score level.

Table 2: Domains and dimensions of FFT used in the MET study

Framework for Teaching (FFT)			
Domain 2: Classroom Environment	Creating an environment of respect and rapport	Domain 3: Instruction	Communicating with students
	Establishing a culture for learning		Using questioning and discussion techniques
	Managing classroom procedures		Engaging students in learning
	Managing student behavior		Using assessment in instruction

In the MET study, raters used the scoring rubrics to rate the first 15 minutes and the 25 to 35 minutes of a lesson on a four-point scale for each dimension. The version of the FFT instrument used in the MET study differs from the most current version (Danielson, 2013) in that the latest version adds two more dimensions—Organizing Physical Space and Demonstrating Flexibility and Responsiveness to Classroom Environment and Instruction—in these two domains respectively.

2.3.1.2. Classroom Assessment Scoring System (CLASS)

CLASS (Pianta, La Paro, & Hamre, 2008) was built on early childhood and elementary classroom research, with a focus on teacher-student interactions that support students' social and academic development. The teacher-student interactions are organized into three domains: 1)

Emotional Support, 2) Classroom Organization, and 3) Instructional Support. Emotional Support domain features dimensions that measure the emotional environment of the classroom; Classroom Organization refers to the ways that teacher structure the classroom processes to manage student behavior and time on learning; and Instructional Support measure along four dimensions to capture teacher supporting students to development conceptual understanding and problem solving skills. Each domain contains several dimensions that are on 7-point scale, and *Student Engagement* score was rated separately from the three domains as a single scoring dimension in its own domain. The version of the CLASS protocol used in K-3 differs slightly in the Instructional Support domain compared to the version used at the upper elementary and the secondary levels, which were used in the MET study (see Table 3).

Table 3: Domains and dimensions of CLASS used in the MET study

Classroom Assessment Scoring System (CLASS)				
Domain	Emotional Support	Classroom Organization	Instructional Support	Student Engagement
Component within Domain	Positive Climate	Behavior Management	Content Understanding	
	Negative Climate	Productivity	Analysis and Problem Solving	
	Teacher Sensitivity	Instructional Learning Formats	Qualify of Feedback	
	Regard for Student Perspective		Instructional Dialogue	

In the MET study, lessons were rated both at the domain level as well as at the sub-domain level, which are the dimensions. The scores that a lesson received at the domain level are just the simple average of all pertaining dimensional scores (except for the dimension of Negative Climate, whose scale is reversed when used to calculate the domain level score).

2.3.1.3 Mathematical Quality of Instruction (MQI)

MQI is a subject-specific protocol that was developed for mathematics instruction with a focus on the richness and rigor of the mathematical content available to students, and the opportunities for mathematical practices during instruction (Hill et al., 2008). The hypothesis in which MQI is grounded is that mathematical work happening in the classroom is distinct from classroom climate and generic classroom strategies. The 3-point version of MQI used in the MET study, referred to as the MQI-Lite, was modified to include only 7 dimensions (see Table 4). It also differs in the level of details at the subscale level from the most current 4-point version (Hill et al., 2012), which is referred to as the MQI-Full by its developers. In accommodation to the policy climate, the *Student Participating in Meaning Making & Reasoning* dimension has been modified to a dimension called *Common Core Aligned Student Practices* with an additional subscale on working with contextualized problems in the latest version. Even though MQI could receive sub-dimensional scores under each dimension, the MET study did not utilize the sub-scales at all. Lessons in the study were rated at a 7.5 minutes' interval using only the dimensions in Table 4, and a holistic score at the dimension level for all four intervals, totaling 30 minutes.

Table 4: Dimensions of MQI used in the MET study

Mathematical Quality of Instruction (MQI)

	Scores at Holistic and Segment Levels	Scores Only at Holistic Level
Dimensions	Richness	Overall mathematical quality of instruction
	Error & Imprecision	Lesson based guess for Mathematical Knowledge for Teaching score
	Explicitness & Thoroughness	
	Student Participation in Meaning Making & Reasoning	
	Working with Students & Mathematics	

2.3.1.4. Protocol for Language Art Observation (PLATO)

PLATO is a subject-specific protocol that was developed for elementary and secondary English Language Arts instruction by combining the use of several dimensions of the CLASS protocol. The version of PLATO used in the MET study includes 6 dimensions of instructions on a 4-point scale, including:

- Intellectual Challenges;
- Classroom Discourse;
- Modeling;
- Strategy Use and Instruction;
- Time Management;
- Behavior Management.

There are also binary content domain scores to indicate the subject areas of the segment to see whether the main content for that time period is reading, writing, literature, or grammar/mechanics. Additionally, there is a binary dimension score for Representation of Content to indicate whether the segment is on ELA or not. Raters scored in every 15 minutes of the lessons and rated two segments in total.

In the following section, validity and reliability studies involving these four protocols are presented to describe the state-of-the-field for the validation program of the teacher quality construct and its measurements.

2.3.2. Validity and Reliability Studies of Observational Protocols

Past research studies, many of which were conducted by the instrument developers themselves, have addressed the reliability and validity issues of these four observational protocols across a variety of contexts (e.g., Grossman et al., 2013; Hamre, Pianta, Mashburn, &

Downer, 2007; Hill, Kapitula, & Umland, 2011; Meyer, Cash, & Mashburn, 2011; Milanowski, 2011). In a study outside of the MET project, Hill and colleagues (2012) emphasized the role of reliability in the design of an observational system—MQI. They contended that the MQI instrument is only a component of the larger comprehensive observational system. The system should include many other components such as the training of raters, systems to prevent rater drifts, and the score aggregation method. They described their rater training processes using the case of the MQI instrument. They also described how they eliminated raters who were out-of-alignment consistently as a way to maintain reliability. Also, they have discussed the consistency across observations and argued that in their case, it was sufficient to get reliable measures by collecting three observations per teacher and assigning two raters per observation.

As Hill and colleagues cautioned against the generalization of their decision rules in other contexts with different protocols, Ho and Kane (2013) also echoed the conclusion that multiple observations per teacher and multiple raters per observation increase reliability measures. They also examined how different combinations of raters and types of observations affect reliability, including employing external and internal raters, observing partial lessons versus the full lessons, and whether or not lessons are chosen by teachers themselves to have evaluators coming in to observe, etc.

Along with the efforts to address the consistency across raters and the consistency across observations, a small number of other studies have tackled the reliability issues on the methodology to collect classroom data. The reliability issues discussed include the processes of adapting observational protocols for large-scale classroom research at elementary grades (Salloum et al., 2016); the consistency across live versus video observational scores (Casabianca et al., 2013); and how raters' fieldnotes are systematically different when using distinct

observational protocols (Bell, Drake, Wilson, Fraiser, & Kim, 2015). Overall, these studies provide new insights into the reliability issues of using observational protocols for the purpose of classroom research.

A series of studies resulted from the MET project have addressed similar types of reliability and the efforts to ensure it within the MET data specifically for the purpose of teacher evaluation. The types of reliability addressed in these studies mostly focus on the reliability issues concerning the role played by people and the external environment, such as the number of raters and observations needed to get to a certain threshold of reliability, rater biases, etc. *Ensuring Fair and Reliable Measures*, a report by Bill & Melinda Gates Foundation (2013), discussed the number of raters and observations desirable to get a reliable measure of teacher performance. For teacher evaluation systems, the report recommends that at least two observations are needed and each should be assessed by a different certified rater. Such recommendations also go into the video data collection and scoring design of the MET study. Further the discussions on what to score by whom in order to ensure a reliable measure of teacher performance, Joe, McClellan, and Holtzman (2014) provided empirical evidence on the high association between scores from certain segments of a lesson and the full one. The authors also examined the cognitive load on raters to use the scoring rubrics of all four instruments in the MET study. They argued that the first thirty minutes of a lesson are necessary and sufficient to represent the full lesson in general, and raters can only focus on a smaller set of dimensions/traits in the scoring rubrics to get the desirable inter-rater reliability. Both practices were hence adopted in the MET study design as well, as only segments up to the first thirty minutes were rated to attain observational ratings of teachers, and different raters were assigned to rate different dimensions within a particular instrument.

Based on the these efforts that go into the design of the MET study to ensure reliability, Park, Chen, and Holtzman (2014) delineated the process to train and monitor raters, and checked for the characteristics of raters, teachers, and classrooms to examine the potential biases that might influence the reliability of the observational scores. The potential biases they investigated include: rater level factors such as raters' background and experience, and their perceptions and training experience collected through survey data; teacher level factors such as gender, years of experience, and racial information; and classroom level factors such as student racial composition and social economic status. They concluded that with rigorous procedures of implementing the observation scoring systems, none of the characteristics (rater, teacher, and classroom) they examined are significantly associated with the scoring accuracy. They did not, however, provide evidence of the potential biases associated with the instrument choice, nor with the content of the lessons being observed. Nonetheless, these three studies above with the MET data have built the foundation for my analyses: One can only further discuss other reliability and validity issues of measures of teacher quality given that raters score reliably to attain the observational scores used in this study, and the observational data rated are sufficient to make inferences with regard to teachers and their quality.

Studies about the validity of the inferences researchers make with regard to teachers using observational instruments are less diverse. In many cases, the implicit inference they try to make is that higher scores attained from the observational measures are associated with higher scores to signify better student performance assessed by a student learning outcome measure. In particular, studies on the validity issues using the MET data include: correlating the observational measure of FFT to the Tripod 7Cs – which is a student survey (Ferguson & Danielson, 2014), and correlating all four protocols to each other, and individually to student

achievement data (Kane, et al., 2012; Walkington & Marder, 2014). In summary, the four observational instruments all show positive correlations with each other. They also demonstrate moderate association with student achievement data, as well with some other student learning outcome measures like the student survey data. But the inferences resulting from using the measurements are not explicitly linked to the validation program of the teacher quality construct for teacher evaluation.

Overall, the analyses that were done by these researchers usually looked at how the simple average of sub-dimensions or individual sub-dimensions can predict the student achievement when examining the validity of an inference one makes using the observational measures. The procedures as well as inferences from these practices, however, are riddled with many issues. First, not all researchers provided rationale for using a particular way to calculate a composite score for a multi-dimensional instrument. Most of the time the researchers just took the simple average across all dimensions. In fact, since the dimensions are inter-correlated, statistical analyses maybe more appropriate to explore the meaning of certain ways to come up with a composite score for a particular instrument on a particular data set. Second, as pointed out by Walkington and Marder (2014) in response to the Gates report on observational measures (Kane et al., 2012), using only the overall score of a teacher rated by observation protocols provides no information on whether any dimensions within each observational instrument has stronger relationship with students' value-added scores, and whether differences for each observation score level were statistically significant. In addition, there is no comparison and contrast among various observational protocols to check consistency across existing generic and subject-specific measures and different ways to aggregate scores. Detailed analysis of observational protocols in context provides more information about teacher quality they are set

to measure, and thus support the inferences that we are making based on the measurement results. Lastly, despite the fact that there are generic and subject-specific observational assessments to examine teacher quality for teacher evaluation, the content of the lessons measured by the assessments is given as though it is an invariant component in the teacher quality measures. My main hypothesis of this study is that content mediates with teaching practices, and hence manifest in the teacher quality measures. Next, I turn to the literature to situate this conjecture to justify the need to attend to the role played by content in getting teacher quality measures reliably and uses them for inferences to evaluate teachers validly.

2.4. Literature Review On Differences in Practices Across Content and Context

In this part of the literature review, I describe past research on how content and context mediate with teaching practices. This study is based on the assumption that teachers' teaching practices are subject to the content and context in which they teach, and differentiated practices may result in differentiated scores attained from observational instruments. To-date, there is a lack of studies to examine the latter part of the assumption across a variety of instruments, especially in the area of mathematics. By providing evidence on the first part of the assumption from the literature, I can support my argument that it is necessary to pay attention to the understudied factors mentioned above in order to examine the second part of the assumption.

2.4.1. Content-specific Practices

Many studies have contributed to the understanding of the influence of subject matter in teaching activities. Through the socio-cultural lens, Grossman and Stodolsky (1995) conducted surveys and interviews with high school teachers to illustrate the salient features of subject subcultures. They argued that teachers of different subjects have dissimilar norms, beliefs, and perceptions of the subject, the curriculum, and their professional community, and thus they

differed in their curricular activities as the subject culture interacts with their teaching practices. They also noted that some subjects, like science and social studies, include a number of different subject areas. Accordingly, there are also many variations in the norms of the subcultures and the beliefs about teaching and learning within areas of those disciplines. Built on the lens of seeing subject matter as the context in which teachers live on a daily basis, Stodolsky and Grossman (1995) compared the conceptions of subject matter and curricular activities of English, social studies, science, mathematics, and foreign language teachers. They found that teachers differed on three features—defined, sequential, and static— in their perceptions of the subjects. In the study, mathematics teachers were more likely to see their disciplines as a well-defined body of knowledge and skills, as more sequential, and as more static than English teachers when talking about their disciplines. Accordingly, mathematics teachers reported more coordination with colleges and more press for coverage of topics during teaching activities other than their beliefs and conceptions of the subject matter. Such findings suggest that teaching activities differ when teaching various subject matter, not only because of the inherent differences in the disciplines themselves, but also because of the subject (sub)cultures in the school in which teachers reside.

Cohen (2013) further explored the content-specific practices and generic practices across two common subjects in the K-8 classrooms—mathematics and ELA, and how the teaching practices got captured by an observational protocol using the video data in MET. She adapted the ELA-specific protocol PLATO to be used in both mathematics and ELA lessons, and focused on the quantitative and qualitative aspects of three practices that are considered widely-used in both subjects: modeling, strategy instruction, and orchestrating discussion. She found that in general teachers did not demonstrate the same instructional practices when teaching different subjects, even when they have demonstrated their strong ability to use the examined practices in one

context. There were significantly more uses of modeling in mathematics than in ELA, but modeling in mathematics was also accompanied by a procedural strategy instruction rather than the conceptual one. Moreover, even when the descriptive statistics show that teachers orchestrated classroom discussion similarly in mathematics and ELA, a qualitative analysis of these discourse moves reveal the non-negligible differences in the nature of these practices.

Research has shown that the content areas of ELA also contribute to the variations in ELA teachers' teaching activities, and thus are reflected by their differentiated observational scores. Grossman, Cohen, and Brown (2014) provided empirical evidence for this type of variation in ELA lessons by examining the observational scores of PLATO. The PLATO rubrics have indicators of content domain, so the rubrics specify whether the lesson segment coded is reading/literature, writing, grammar/mechanics, or vocabulary. These authors found that teachers' scores on individual dimensions differed significantly for distinct content domains in ELA, despite the fact that the average scores across dimensions were not significantly different between reading/literature and writing. They suggested that there might be consequential variations in terms of the teaching practices that teachers use when teaching different content domains within ELA, and that variations also exist under other contexts, such as grade level and the composition of the classroom.

Among the areas of mathematics, there is a lack of studies on teaching practices in mathematics that highlight the role played by its subject areas. Building on Ball and colleagues' work on Mathematical Knowledge for Teaching (MKT), researchers have been investigating the specific knowledge for teaching situated in a particular area of mathematics, such as Algebra (McCrory, Floden, Ferrini-Mundy, Reckase, & Senk, 2012) and Geometry (Herbst & Kosko, 2014). The focus for this line of research is not to highlight the subject-area-specific practices in

order to compare and contrast them, but rather to provide frameworks to conceptualize knowledge using teaching practices grounded in particular content. Statistics, however, seems to be the exception. Research on statistical knowledge for teaching always state on the forefront that there are fundamental differences in knowledge and practices for teaching statistics in comparison to other areas of mathematics. This claim is rooted in the shared belief among current statistics educators that while probability can be considered as a field of mathematics, statistics is a mathematical science that is a different discipline from mathematics (Cobb & Moore, 1997; Moore, 1992). In particular, the exploratory data analysis that is taught at the middle school level as designated by the Common Core State Standards (CCSS) originated from empirical science and the need to handle data rather than from theorems and axioms (Tukey, 1977). Accordingly, statistics educators argue that the content knowledge and the pedagogical content knowledge needed to teach statistics distinguish from mathematical knowledge for teaching (Burgess, 2007; Cobb & Moore, 1997; Groth, 2007). Even experienced mathematics teachers may not be able to teach statistics well because they are not familiar with the norms and cultures in the field of statistics, and they may not be aware of the differences between mathematical thinking and statistical thinking (Sanchez & Blancarte, 2008). Still, these existing studies focus more on the teaching practices across areas of mathematics from a theoretical standpoint, as they provide evidence from instruction to conceptualize the knowledge for teaching these subject areas. The questions that remain to be answered are: Do conceptually different teacher knowledge and practices also get captured by the generic or the math-specific observational instruments? Do the differences in practices and knowledge result in differentiated observational ratings? A comparable study to what Grossman et al. and Cohen have done in light of other observational instruments and in mathematics can further the insights of these authors.

In sum, the two recent studies regarding PLATO documented above show efforts to connect content-specific practices to their manifestations in observational scores as measured by one particular subject-specific instrument in ELA. This study can build on what these researchers have done and contribute to this body of literature in two ways. First, I examine whether there are differentiated observational scores across subject areas of mathematics, as measured by both generic instruments and math-specific instrument. Also, the efforts so far have been focusing on only PLATO and the adapted version of PLATO, which is an ELA-specific instrument. Whether other instruments, including both the generic and math-specific instruments, demonstrate similar results in capturing differences between teachers' practices in ELA and in mathematics is still not investigated.

2.4.2. Context-specific Practices

There are many other contextual factors that may contribute to the differences in teaching practices and the differentiated observational scores attained from instruments. Kennedy (2010) suggested that observers of teachers tend to make the fundamental attribution error (Gilbert & Malone, 1995; Humphrey, 1985; Ross, 1977), and attribute teachers' own personal characteristics to his or her teaching practices. She argued that there are many situational factors in play: the amount of the time for teachers to plan lessons, the curriculum that he or she has to follow, or some other school and district requirements. Accordingly, it is teachers' personal characteristics together with the situational characteristics that influence their teaching practices, and as a result influence student learning outcomes. It is important to be mindful that an overemphasis of teacher quality in individuals might not be able to account for everything that teachers do and how their students perform.

Other research of the MET study has shown that there are non-negligible variations in teachers' observational scores attained from different observational instruments across grade levels (Mihaly & McCaffrey, 2014). Generally, teachers in Grade 4 and 5 have significantly higher simple average scores than teachers of higher grades in CLASS, FFT and PLATO, and also higher dimensional scores within specific instruments. These authors have cautioned the principals and other policymakers to consider this inherent trend of the observational measures, and to take that into consideration when targeting professional development resources and making low-stakes or high-stakes decisions regarding teachers.

Overall, content and context contribute to the differences in teaching practices, and such differences have been shown to manifest in the observational scores in some content areas and with some instruments. But how that relates to the validation program of the teacher quality construct in the context of teacher evaluation is a missing component, especially for the subject of mathematics.

2.5. Argument-based Approach to the Validation of Teacher Quality

This section focuses on the guiding framework to examine the construct of teacher quality in the context of teacher evaluation, and the types of validity and reliability that this study focuses on based on the gaps identified in the previous sections of the literature review.

Construct validity is an indispensable property of any measurement. According to Cronbach and Meehl (1955), who greatly developed the concept of construct validity, there should be a strong theoretical foundation or a nomological network behind the construct being measured. Moreover, the validation of the construct is about validating the proposed interpretations and inferences that one makes based on the test scores. Rather than seeking only evidence to confirm one's preconception of the proposed interpretations, alternative

interpretations should also be evaluated. Thus, validation requires a research program to evaluate the measurement from many different aspects, and gauge against alternative interpretations.

Although many researchers have felt that it is rare for most constructs, especially in social science, to have a strong theory or a clear nomological network behind the construct as suggested by Cronbach and Meehl, the principles of construct validation put forward by these two researchers still hold. It is just a matter of whether the program is a strong program (validating the construct by stating the theory and devise challenges to the theory) or a weak one (validating the construct by providing descriptions of correlation to any other variables).

Built on these principles, Kane (2012) extended the research program notion of the construct validity, and suggested that researchers should not only lay out "the *interpretative argument* that explicitly states the claims being made, as a chain or network of inferences and assumptions leading from the observed assessment performances to the interpretation and use of the test scores (p. 68)," but also include the evaluation of the consequences/decision procedures in the research program of validation. He referred to this as an argument-based approach. Under his framework, to make claims about the interpretations of measuring teacher quality using observational protocols, researchers need to tackle questions of domain coverage, reliability, and the potential sources of bias. Moreover, to make claims about the consequences of measuring teacher quality using observational protocols, researchers and policymakers face additional questions related to appropriateness, relevance, and fairness of the measurements for their intended purposes, and they also need to provide more empirical grounds to address these questions.

In this study, I contribute to the validation program of the teacher quality construct by examining properties of several well-established instruments, and potential biases towards the

inferences from using their observational ratings in teacher evaluation. I used the classical definitions of validity and reliability where reliability is defined as the within-teacher consistency of the measures, and validity is defined as the ability of the measure to capture the underlying construct of teacher quality (Brennan, 2006; Kane, 2006, 2013). I do not, however, exhaust all types of within-teacher consistency in this study. I focus on the internal consistency of the instrument to rate the same teachers' instruction across content, and across different ways to attain the observational ratings for use. I also focus on the external consistency between two instruments to rate the same teachers' mathematical instruction on the same set of lessons. If the same teacher consistently receives higher ratings from teaching one subject over the other, or from one particular observational protocol, there are potential biases in terms of the content of the lessons that he or she is observed teaching, and there is the problem of inappropriateness in using one observational protocol to score across lessons of different content. Most importantly in terms of teachers' interests, it is unfair to those who are observed in teaching the content that on average receives lower ratings so that their evaluation results do not look prominent as they could be. Other types of within-teacher consistency, such as the inter-temporal consistency of the measures over time (that is, measures of teacher quality between a teacher in Year One and the same teacher in Year Two) is not addressed in this study. Validity issues regarding these four instruments are intertwined with reliability issues in examining the potentially influential factors towards teachers' observational ratings, as well as the interpretations and consequences of using the scores. During the process of examining internal consistency within teachers, by relating the two measures of teacher quality as manifested in different content areas from the same instrument, I examine the ability of the instrument to capture a unifying teacher quality (or a unifying aspect of teacher quality) across content. Similarly, while examining the external

consistency between instruments, by relating the two measures of teacher quality in the same content from different instruments, I add to the validity argument that both instruments are similar/dissimilar in their ability to measure teacher quality (or particular aspects of teacher quality).

In sum, there are three sources of potential bias towards implementing the observation component in the context of teacher evaluation: The bias of the content, the bias of the instrument choice, and the bias of the score aggregation methods. It is of great importance to examine these factors and provide evidence on the validity and reliability of the inferences one makes with regard to the use of the scores from each instrument, and discuss the consequences of getting observational ratings under different contexts.

2.6. Conceptual Framework

In this section, I demonstrate the conceptual framework that ties all the components discussed above in the literature review to situate the study. The underlying assumption for most current teacher evaluation practices is that there is a "true" teacher quality possessed by individual teachers. Sufficient samples of their performances, qualifications, and effectiveness and reasonable measurements of these samples should inform educators on how much quality the teachers most likely have at the moment of the assessments. Even though teacher quality is a developing trait of teachers, but generally it is also considered as a relatively constant characteristic for a range of time (e.g., a school year). So the interpretations of the scores resulting from those assessments and decisions made using those scores regarding teachers during a particular time frame are considered as legitimate actions. Under current policy environment, evaluating teachers and using evaluation results for decisions about tenure, retention, compensation, and resource targeting are educational priorities in many states. Even

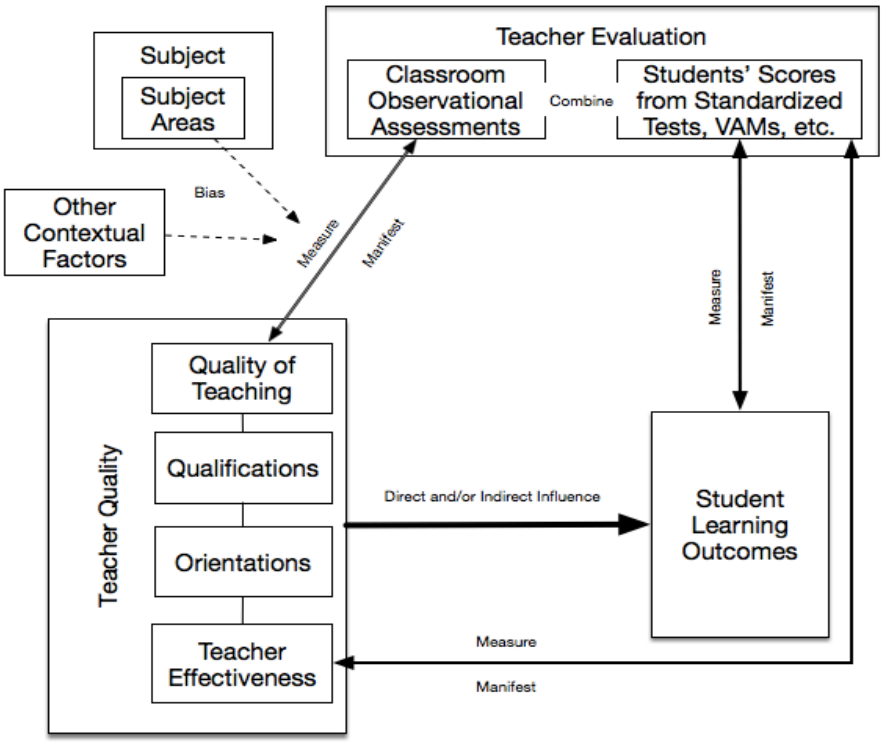
with states who do not write the teacher evaluation officially into bills, the usage of observations still occurs at the level of schools and districts. There is some criticism about such annual evaluation because scholars concern that the history of the teachers as well as the history of their incoming students are not accounted for. Some remedies to these criticisms include using teacher value-added measures, teachers' career-to-date performance (Staiger & Kane, 2014), and multiple measures for each aspect of teacher quality (Kane et al., 2012).

The suggestion of using multiple measures is based on the assumption that each of the measure validly and reliably reflects the "true" teacher quality so that we are confident about the scores we get in order to use them for inferences and decision-making. For classroom observational assessments that are designed to measure complex teaching activities rather than simple teacher background questions such as degree obtained, we need to take into consideration all potential biases inherited in the scores and how ways of using the scores render different interpretations and consequences. Based on the literature, I hypothesize that one main source of bias results from the content of teaching. In other words, teaching activities not only vary when teachers are teaching different subjects, but also vary in different subject areas within mathematics. Observational protocols are used to measure teaching practices, so there may be some systematic differences in the observational scores across lessons of various content, even when the lessons are taught by the same teacher. Other contextual factors may also contribute to the systematic differences, if any, in the ratings of different content, but they may not be able to explain all the variance. Observational protocols may not be able to capture the nuances in teaching diverse content, and hence render significantly and practically different ratings and evaluation results for teachers. Such difference is not alarming if small, or not around the cut-off points to differentiate teachers. But if not, states who are building teacher evaluation systems

should think about the nature of the system before they use the scores to fire or reward teachers, and adapt the system to account for the potential biases. Lastly, what observational protocol to use is usually decided at the state or district level. If different types of observational protocols do not give consistent estimates of teacher quality, the stakeholders must be aware of the consequences, and support their own rationale of choosing a particular protocol based on their educational agenda locally and state-wise.

Based on these considerations, the diagram below summarizes the relationships among teacher quality, teacher evaluation, and potential biases that may influence the use of classroom protocols to measures teacher quality for teacher evaluation. Overall, the framework embodies what Kane referred to as the teacher quality validation program, and how this study contributes to the program.

Figure 1: Conceptual framework for the construct validation program and teacher evaluation



CHAPTER 3

MEASURES AND SAMPLES

3.1. MET Project Data Overview

The sample in this study is drawn from Measures of Effective Teaching (MET) project supported by the Bill and Melinda Gates Foundation (2010). This project is a large-scale study of teacher quality and teacher effectiveness featuring near 3,000 teachers of Grades 4-9 in six school districts—New York City, Dallas, Denver, Charlotte-Mecklenburg, Memphis, and Hillsborough County (Florida). In the Spring semester of school year 2009-2010 (Year One), data were collected in 2,741 participating teachers' classrooms. In the school year of 2010-2011 (Year Two), 2,086 teachers remained in the study after attribution, and data were collected on them together with new participants for another year. Among the teachers who remained, 1,559 were randomly assigned to sections of students within schools that were assembled by their principals, while the rest was labeled as the non-randomized sample due to non-compliance. Even though the assignment of teachers to students are not completely random as these teachers still taught students with the same demographic makeup in their schools, the MET project has made an effort to control for the student effect on teachers' instructional quality.

The MET study measured teachers' quality by training raters to score videos of teachers at scale[4], rather than to have raters coming to teachers' classrooms to do live observations. The project collected over 25,000 videos, and 11,500 of them are available for online streaming. All videos from the randomized teacher sample were scored in the two generic observational instruments (FFT and CLASS), and one of the subject specific instruments—MQI, PLATO, or

[4] For details on the rater training, assignment, and calibration processes for the video scoring, please refer to the section 2.3.2 for the literature review of several studies on the MET project to ensure reliability.

Quality Science Teaching (QST)—depending on whether they are mathematics, ELA, or science lessons. Additionally, a sample of 2,000 videos collected from Year One with complete data was chosen as the Plan B data. These videos, regardless whether they were collected from teachers in the randomized blocks, were scored in not only the two generic instruments and one subject-specific instrument, but also the UTeach Observational Protocol (UTOP) if they are mathematics or science lessons. In total, about 60% of all videos were scored[5]. Among them, videos of randomized ELA and mathematics teachers plus the Plan B videos were scored in FFT and CLASS, videos of randomized ELA teachers plus the Plan B ELA lessons were scored in PLATO, and videos of randomized math teachers plus the Plan B mathematics videos were scored in MQI. Table 5 below summarizes the teacher population that is relevant for this study.

Table 5: Numbers of generalist teachers, ELA teachers, and mathematics teacher by year

	Year One	Year Two	
Generalist Teachers (who taught both ELA and mathematics in Grades 4-6)	652	365	Randomized: 309
			Non-Randomized: 56
Mathematics Teachers (including both generalist and math-specialist teachers in Grades 4-9)	1,515	1,025	Randomized: 774
			Non-Randomized: 251
ELA Teachers (including both generalist and ELA-specialist teachers in Grades 4-9)	1,396	1,079	Randomized: 807
			Non-randomized: 272
Total	2,501	1,739	Randomized: 1,272
			Non-randomized: 467

Note: Cells in the Total row do not equal to the sum of the column because the row categories are overlapped. The mathematics teachers sample include those who are generalist teachers who also taught mathematics, as well as specialist teachers who only taught mathematics at the time of the data collection.

[5] This information was gathered in the MET project workshop at AERA 2015.

For this study, the generalist teachers in Grades 4-6 whose videos were scored make up the sample to answer the both parts of the first research question. The mathematics teachers in Grades 4-9 (including those who were generalist teachers that also taught mathematics, as well as those who were specialist teachers that only taught mathematics at the time of the data collection) whose videos were scored make up the sample to answer the second and third research questions. In order to identify the subsamples within the mathematics teacher population to answer the research question regarding observational scores' differences across areas of mathematics, I first categorized mathematics lessons into four main subject areas. The next section focuses on the effort of categorization using lessons' focal topics information in the MET data.

3.2. The Mapping of Focal Topics onto Mathematical Subject Areas

In the MET data set, there are 9,728 mathematics lessons in total. Among them, 3,898 lessons are labeled as *Random Topic,* while the rest are labeled by various focal topics in mathematics prescribed by the MET researchers beforehand. In order to achieve variety and diversity in content, participating teachers chose from a list of focal topics to record their teaching of that topic for at least half of the videos collected, and chose topics of their choice for the rest of the videos collected.

For each of the focal topic prescribed by the MET study, I found the specific standards related to it and the specific domains that such standards belong to in the Common Core State Standards of Mathematics (CCSS, 2010). By mapping the focal topics to the CCSS domains, I was able to group the focal topics into three areas of mathematics, including Numbers & Operations, Geometry, and Algebra & Algebraic Thinking. Table 6 below shows the list of focal

topics, the related standards for each focal topic, the domain that the standards belong to, and the identified subject area based on the information.

Table 6: Mapping of focal topics to subject areas within mathematics according to CCSS

Focal Topics	Frequency	Common Core Domain	Grade Level in Common Core	Subject Area
Adding and subtracting fractions	519	Numbers & Operations-Fractions	5.NF.A.1	Numbers & Operations
Completing function tables and finding function rules	25	Functions	8.F.A.1; 8.F.A.2	Algebra & Algebraic Thinking
Creating, analyzing tables, graphs and equations to describe linear functions and other relationships	44	Functions	8.F.B.4	Algebra & Algebraic Thinking
Decimals and their meaning; relationship of decimals to fractions	407	Numbers & Operations-Fractions	4.NF.C.6	Numbers & Operations
Determining the area and perimeter of two-dimensional shapes	383	Geometry	6. G.A.1	Geometry
Exponents & Exponential Functions	45	High school: Functions	HSF.LE.A	Algebra & Algebraic Thinking
Functions and Pythagorean Theorem	41	Geometry	8.G.B.7	Geometry
Functions or polynomials	70	High school: Algebra	HSA.APR.A.1; HSA.APR.B.2; HSA.APR.B.3; HSA.APR.C.4	Algebra & Algebraic Thinking
Graphing linear equations	204	High school: Algebra	HSA.CED.A.2	Algebra & Algebraic Thinking
Linear equations	113	High school: Algebra	HSA.CED.A	Algebra & Algebraic Thinking
Multi-digit multiplication and division	954	The Number System	6. NS.B.2	Numbers & Operations

Table 6 (cont'd)

Focal Topics	Frequency	Common Core Domain	Grade Level in Common Core	Subject Area
Multiplication and division of fractions or decimals	181	Numbers & Operations-Fractions	5.NF.B.4; 5.NF.B.7; 6.NS.B.3	Numbers & Operations
operations on rational numbers	154	The Number System	7.NS.A.1; 7.NS.A.2	Numbers & Operations
Operations with negative integers	89	The Number System	6.NS.C.6	Numbers & Operations
Percents and operations involving percents	166	Ratios & Proportional Relationships	6. RPA.3.C	Algebra & Algebraic Thinking
Polynomials & Factoring	84	High school: Algebra	HSA.APR.A.1; HSA.APR.B.2	Algebra & Algebraic Thinking
Quadratic Equations & Functions	60	High school: Functions; High school: Algebra	HSA.CED.A.1	Algebra & Algebraic Thinking
Random Topic	3898			Indeterminable
Ratio, rate, and proportional reasoning	572	Ratios & Proportional Relationships	6. RPA.1; 6. RPA.2; 6.RPA.3	Algebra & Algebraic Thinking
Ratio, rate, proportional reasoning, and percent				Algebra & Algebraic Thinking
Rational algebraic expressions, equations, and functions	61	Expressions & Equations; Function	8.EE.C.7.B	Algebra & Algebraic Thinking
Representing and solving linear functions and linear equations	334	Functions; Expressions & Equations	8.F.A.12; 8.EE.C.7	Algebra & Algebraic Thinking
Representing patterns, models, and relationships (e.g., story problems) as simple equations	402	Expressions & Equations; Operations & Algebraic Thinking	6.EE.C.9; 5.OA.B.3	Algebra & Algebraic Thinking
Simplifying expressions and solving linear equations	239	Expressions & Equations	6.EE.A.3; 6.EE.B.5	Algebra & Algebraic Thinking

Table 6 (cont'd)

Focal Topics	Frequency	Common Core Domain	Grade Level in Common Core	Subject Area
Solving addition and subtraction problems involving integers	60	The Number System	6.NS.A.1; 6.NS.B	Numbers & Operations
Solving and graph two step equations and inequalities	61	Expressions & Equations	7.EE.B.4	Algebra & Algebraic Thinking
Solving multiplication and division equations involving integers	37	Expressions & Equations	6.EE.A.1	Algebra & Algebraic Thinking
Solving systems of linear equations	217	High school: Algebra	HSA.REI.C.6	Algebra & Algebraic Thinking
Using the commutative, associative, identity and distributive properties	26	Expression & Equations	6.EE.A.3	Algebra & Algebraic Thinking
Writing, interpreting, and/or using mathematical expressions and equations	199	Operations & Algebraic Thinking; Expressions & Equations; High school: Algebra	5.OA.A.1; 6.EE.C.9; 6.EE. B.6; HSA.SSE.A.1	Algebra & Algebraic Thinking
Total	9,728			

After grouping the lessons by focal topics, there are 2,530 lessons in the subject area of Numbers & Operations, 424 lessons in Geometry, and 2,793 lessons in Algebra & Algebraic Thinking. Moreover, based on CCSS, I identified another important subject area in the current mathematics curriculum—Statistics & Probability—that is not readily identifiable by given focal topics. The lessons in this subject area are generally but not always labeled as Random Topic, as there are some lessons mislabeled by teachers themselves. In order to find sufficient Statistics & Probability lessons, I used videos available for online streaming to locate lessons that fall into this subject area. I first narrowed down the lessons to be in Grade 6 to 8, because this is the grade band in which Statistics & Probability are more salient in the curriculum as recommended by the

CCSS standards. I went through all the available videos in this grade band, spent at least 1-2 minutes to watch the lesson until I could identify the topics taught by the teachers, and confirmed whether it was a lesson in Statistics & Probability or not. After skimming through all possible candidate videos, I identified 84 statistics lessons with content ranging from probability (theoretical probability, experimental probability, and probability of complex events, etc.), descriptive statistics, and statistical representations. 73 of them are labeled as "Random Topic" in the video information file, while nine of them are labeled as some other focal topics, such as "Creating, analyzing tables, graphs, and equations to describe linear functions and other relationships", and "Ratio, rate, and proportional reasoning". Moreover, there are two lessons that I found in the video database that are not present in the video information file and the observational scores files.

To clean the data, first, I watched the videos with other focal topics to see if they are really statistical rather than mathematical by examining the content taught. After re-watching the lessons, I confirmed that these nine lessons are indeed Statistics & Probability lessons, and hence changed their focal topics in the video information files. Then I replaced focal topics of the statistics lessons with the focal topics that I defined in all relevant files to include these lessons' observational scores for analyses. In total, there are 33 Statistics & Probability lessons in Year One, and 49 lessons in Year Two, with two other videos available online but missing in the video information file and item-level observational score files. The list of videos for these Statistics & Probability lessons and their content by grade level can be found in Appendix A.

Based on the identified subject areas within mathematics, the numbers of mathematics teachers who teach either two of the subject areas are as follows:

- Numbers & Operations and Algebra & Algebraic Thinking: 406 teachers

- Numbers & Operations and Geometry: 219 teachers
- Numbers & Operations and Statistics & Probability: 37 teachers
- Algebra & Algebraic Thinking and Geometry: 38 teachers
- Algebra & Algebraic Thinking and Statistics & Probability: 56 teachers.
- Geometry and Statistics & Probability: 5 teachers.

A small sample size may result in insufficient statistical power to detect any real effect in statistical tests. Accordingly, the pairs of subject areas that are included in the study are: 1) Numbers & Operations (NO) and Algebra & Algebraic Thinking (AA); 2) Numbers & Operations (NO) and Geometry (G), and 3) Algebra & Algebraic Thinking (AA) and Statistics & Probability (SP). In order to avoid the issue of correlation from repeated measures on the same teachers since there are many returning teachers in the second year, I did the analysis on teachers' observational measures by year. The sample size to answer the first, second and third research questions on difference in observational measures across diverse subject and subject areas within mathematics are summarized in Table 7 below.

Table 7: Sample size for each group of comparisons by year

Research Question	Instruments		Year One	Year Two
Research Question 1—Generalist teachers who taught both ELA and Math in Grades 4-6	FFT (math) vs. FFT (ELA)		440	313
	CLASS (math) vs. CLASS (ELA)		440	313
	PLATO vs. MQI		430	310
Research Question 2—All teachers who taught mathematics in Grades 4-9	FFT vs. CLASS		978	772
	FFT vs. MQI		971	770
	CLASS vs. MQI		971	770
Research Question 3—Mathematics teachers who taught two subject areas within mathematics in Grades 4-9	FFT	NO vs. AA	230	175
		NO vs. G	135	84
		AA vs. SP	56	
	CLASS	NO vs. AA	231	175
		NO vs. G	135	84
		AA vs. SP	56	
	MQI	NO vs. AA	221	171
		NO vs. G	125	81
		AA vs. SP	53	

Noted that I did not conduct the analysis by year for the group of teachers who teach both Algebra & Algebraic Thinking and Statistics & Probability because the majority of them are distinct individuals in the sample. There are two teachers, however, who were represented in both years' samples to teach both of these two subject areas. In order to get sufficient sample size, I had to combined Year One and Year Two subsamples to report; but at the same time, in order to avoid the repeated measure issue, I only included these two teachers' one year's ratings

in the analysis. Specifically, I deleted ratings received in Year One of one teacher and ratings in Year Two of the other teacher randomly.

3.3. Measures

The variables used to answer the research questions are shown in Table 8. These variables include: item level observational scores variables, lesson information variables, and school variables. The names of the variables are not perfectly consistent across different files, so I only provide one variant of the variable names in the class, and provide descriptions of this type of variables.

Table 8: Variables in the MET study to be used in the analyses

Main Variable	Variable Family	Variable Type	Description
Lesson Information	Focal Topic of the lesson (FOCALTOPIC)	String	The focal topic of the lesson indicated by teachers.
	Subject of the lesson (SUBJECT)	String	ELA or Math.
	Video ID (VIDEO_ICPSR_ID)	Nominal	Unique identifier of each lesson recorded.
	Grade Level (GRADE)	Nominal	Grade level of the lesson.
FFT	Creating an Environment of Respect and Rapport (FFT_CERR)	Ordinal	Lesson level score on a 4-point scale. One rating for the union of 0-15 minutes and 25-35 minutes of a lesson. 1 is the lowest score, and 4 is the highest score. When averaged two raters' scores in double scored lessons, half point is possible.
FFT	Establishing a Culture for Learning (FFT_ECL)	Ordinal	Lesson level score on a 4-point scale. One rating for the union of 0-15 minutes and 25-35 minutes of a lesson. 1 is the lowest score, and 4 is the highest score. When averaged two raters' scores in double scored lessons, half point is possible.

Table 8 (cont'd)

Main Variable	Variable Family	Variable Type	Description
FFT	Managing Student Behavior (FFT_MSB)	Ordinal	Lesson level score on a 4-point scale. One rating for the union of 0-15 minutes and 25-35 minutes of a lesson. 1 is the lowest score, and 4 is the highest score. When averaged two raters' scores in double scored lessons, half point is possible.
FFT	Communicating with Students (FFT_CS)	Ordinal	Lesson level score on a 4-point scale. One rating for the union of 0-15 minutes and 25-35 minutes of a lesson. 1 is the lowest score, and 4 is the highest score. When averaged two raters' scores in double scored lessons, half point is possible.
FFT	Using Questioning and Discussion Techniques (FFT_UQDT)	Ordinal	Lesson level score on a 4-point scale. One rating for the union of 0-15 minutes and 25-35 minutes of a lesson. 1 is the lowest score, and 4 is the highest score. When averaged two raters' scores in double scored lessons, half point is possible.
FFT	Engaging Students in Learning (FFT_ESL)	Ordinal	Lesson level score on a 4-point scale. One rating for the union of 0-15 minutes and 25-35 minutes of a lesson. 1 is the lowest score, and 4 is the highest score. When averaged two raters' scores in double scored lessons, half point is possible.
FFT	Using Assessment in Instruction (FFT_UAI)	Ordinal	Lesson level score on a 4-point scale. One rating for the union of 0-15 minutes and 25-35 minutes of a lesson. 1 is the lowest score, and 4 is the highest score. When averaged two raters' scores in double scored lessons, half point is possible.

Table 8 (cont'd)

Main Variable	Variable Family	Variable Type	Description
CLASS	Positive Climate (CLASS_PC)	Ordinal	There are both segment level scores and lesson level scores on a 7-point scale for this variable family. Segment length is 15 minutes and 2 segments were rated. 1 is the lowest score, and 7 is the highest score.
CLASS	Negative Climate (CLASS_NC)	Ordinal	There are both segment level scores and lesson level scores on a 7-point scale for this variable family. Segment length is 15 minutes and 2 segments were rated. 1 is the lowest score, and 7 is the highest score. Noted that this dimension is in the opposite direction in contrast to other dimension: higher ratings indicate lower quality of teaching from the teacher.
CLASS	Teacher Sensitivity (CLASS_TS)	Ordinal	There are both segment level score and lesson level scores on a 7-point scale for this variable family. Segment length is 15 minutes and 2 segments were rated. 1 is the lowest score, and 7 is the highest score.
CLASS	Regard for Student Perspective (CLASS_RSP)	Ordinal	There are both segment level score and lesson level scores on a 7-point scale for this variable family. Segment length is 15 minutes and 2 segments were rated. 1 is the lowest score, and 7 is the highest score.
CLASS	Behavior Management (CLASS_BM)	Ordinal	There are both segment level score and lesson level scores on a 7-point scale for this variable family. Segment length is 15 minutes and 2 segments were rated. 1 is the lowest score, and 7 is the highest score.

Table 8 (cont'd)

Main Variable	Variable Family	Variable Type	Description
CLASS	Productivity (CLASS_PC)	Ordinal	There are both segment level score and lesson level scores on a 7-point scale for this variable family. Segment length is 15 minutes and 2 segments were rated. 1 is the lowest score, and 7 is the highest score.
CLASS	Instructional Learning Format (CLASS_ILF)	Ordinal	There are both segment level score and lesson level scores on a 7-point scale for this variable family. Segment length is 15 minutes and 2 segments were rated. 1 is the lowest score, and 7 is the highest score.
CLASS	Content Understanding (CLASS_CU)	Ordinal	There are both segment level score and lesson level scores on a 7-point scale for this variable family. Segment length is 15 minutes and 2 segments were rated. 1 is the lowest score, and 7 is the highest score.
CLASS	Analysis and Problem Solving (CLASS_APS)	Ordinal	There are both segment level score and lesson level scores on a 7-point scale for this variable family. Segment length is 15 minutes and 2 segments were rated. 1 is the lowest score, and 7 is the highest score.
CLASS	Quality of Feedback (QF)	Ordinal	There are both segment level score and lesson level scores on a 7-point scale for this variable family. Segment length is 15 minutes and 2 segments were rated. 1 is the lowest score, and 7 is the highest score.
CLASS	Instructional Dialogue (CLASS_ID)	Ordinal	There are both segment level score and lesson level scores on a 7-point scale for this variable family. Segment length is 15 minutes and 2 segments were rated. 1 is the lowest score, and 7 is the highest score.

Table 8 (cont'd)

Main Variable	Variable Family	Variable Type	Description
MQI	Richness of Content (MQI_RICH)	Ordinal	There are both segment level scores and lesson level scores on a 3-point scale for this variable family. 1 is the lowest score, 3 is the highest score. When averaging from two raters, half point is possible. Segment length is 7.5 minutes and there are 4 segments in total for each lesson.
MQI	Student Participation in Meaning Making and Reasoning (MQI_SPMMR)	Ordinal	There are both segment level scores and lesson level scores on a 3-point scale for this variable family. 1 is the lowest score, and 3 is the highest score. When averaging from two raters, half point is possible. Segment length is 7.5 minutes and there are 4 segments in total for each lesson.
MQI	Working with Students & Mathematics (MQI_WSM)	Ordinal	There are both segment level scores and lesson level scores on a 3-point scale for this variable family. 1 is the lowest score, and 3 is the highest score. When averaging from two raters, half point is possible. Segment length is 7.5 minutes and there are 4 segments in total for each lesson.
MQI	Error & Imprecision (MQI_EI)	Ordinal	There are both segment level scores and lesson level scores on a 3-point scale for this variable family. 1 is the lowest score, and 3 is the highest score. When averaging from two raters, half point is possible. Segment length is 7.5 minutes and there are 4 segments in total for each lesson. Noted that this variable is in the opposite direction in contrast to other variables. The higher the scores the more or worse the mistake, indicating lower teaching quality.

Table 8 (cont'd)

Main Variable	Variable Family	Variable Type	Description
PLATO	Intellectual Challenge (PLATO_IC)	Ordinal	Segment level scores on a 4-point scale for this variable family. 1 is the lowest score, and 3 is the highest score. Each segment is 15 minutes long. Two segments in total for each lesson were rated.
PLATO	Classroom Discourse (PLATO_CD)	Ordinal	Segment level scores on a 4-point scale for this variable family. 1 is the lowest score, and 3 is the highest score. Each segment is 15 minutes long. Two segments in total for each lesson were rated.
PLATO	Modeling (PLATO_M)	Ordinal	Segment level scores on a 4-point scale for this variable family. 1 is the lowest score, and 3 is the highest score. Each segment is 15 minutes long. Two segments in total for each lesson were rated.
PLATO	Strategy Use and Instruction (PLATO_SUI)	Ordinal	Segment level scores on a 4-point scale for this variable family. 1 is the lowest score, and 3 is the highest score. Each segment is 15 minutes long. Two segments in total for each lesson were rated.
PLATO	Time Management (PLATO_TM)	Ordinal	Segment level scores on a 4-point scale for this variable family. 1 is the lowest score, and 3 is the highest score. Each segment is 15 minutes long. Two segments in total for each lesson were rated.
PLATO	Behavior Management (PLATO_BM)	Ordinal	Segment level scores on a 4-point scale for this variable family. 1 is the lowest score, and 3 is the highest score. Each segment is 15 minutes long. Two segments in total for each lesson were rated.
BACKGROUND	District ID (DISTRICT_ICPSR_ID)	Nominal	The district of the school.

The variables listed above and their variants were used to generate the composite scores at the lesson level first for each video, and then aggregated at the teacher level depending on the

samples for the comparison pairs. In the next chapter, I illustrate two ways of generated composite scores using these dimensional level scores, and how these composite scores are used for the comparisons in the future chapters.

CHAPTER 4

EMPIRICAL APPROACH TO CAPTURE TEACHER QUALITY USING OBSERVATIONAL INSTRUMENTS

As reviewed in Chapter 2, simply averaging the ratings across all dimension of an instrument may be an intuitive thing to do to get a univariate score of teachers, and such approach is generally the practice that most teacher evaluation systems have taken. But this method is not grounded in the design of the instrument and the hypothesized underlying construct captured. Moreover, the meaning of the univariate composite is unclear.

In this study, I used both the simple average algorithm and the Principal Component Analysis (PCA) algorithm to generate two sets of composite scores to represent teachers' quality, and examine teachers' observational ratings attained from both algorithms. With the use of different algorithms to generate composite scores, I also compare whether the common practice of using a simple average would yield different results than the use of factor analysis. If both algorithms show highly consistent results in most of the comparisons I made, especially those with significance detected, even though the meaning of these two sets of composite scores are dissimilar and hence require different interpretations, they did not differ from each other much as a numeric result of the observation component in teacher evaluation. This chapter describes the two algorithms to generate composite scores, and presents a conceptual mapping among instruments to link conceptually matched components across instruments based the meanings and interpretations of the component in PCA. Only those pairs of the components that are conceptually related to a large extent are used for comparisons of teachers' observational scores across instruments and content in the future chapters. This decision is made because it is

meaningless to compare two sub-constructs across instruments and content if they are not measuring the same quality and are supposed to be differentiated in scores.

4.1. Getting Simple Average Composite Scores

Before taking the average of the dimensions, I first identified the meaning of the dimensions in all instruments to make sure they are keyed in the same direction. For those negative-keyed variables—***Negative Climate*** in CLASS, and ***Error & Imprecision*** in MQI—I recoded these variables into their inverses based on their original scales. For example, CLASS dimensions are rated at a 7-point scale, so a score of 1 in ***Negative Climate*** was recoded as the score of 7 in the inverse variable of Negative Climate (NEGATIVE_CLIMATE_INV).

In addition, because there are many repeated measures resulted from double scoring and segment level scoring of the same dimension within each lesson, I needed to average these repeated measures to get a single score for each dimension at the lesson level first. There are two types of aggregation in order to get ratings of teachers at each dimension for each instrument before calculating the simple average composite scores. First, for those instruments that rate lessons at the segment level (CLASS and MQI), the segment ratings of each dimension were aggregated within the lessons first to get segment level aggregated dimension scores. Second, for lessons rated by two raters, the ratings from each rater were also aggregated at each dimension within the instrument before using them for calculating component scores.

In summary, to attain the simple averages of teachers for each instrument, I first reversed the negative-keyed dimensions, then calculated the averages at the lesson level, and finally aggregated at the teacher levels based on different grouping of teachers and subsets of teachers' lessons according to samples needed for comparisons.

4.2. Getting Composite Scores with PCA

Among the four instruments in the MET project, FFT is the most widely studied tool in terms of making sense of its multi-dimensional ratings and of how it can be used for decision-making in evaluating teachers. Research has suggested that the eight dimensions in the two domains used in the MET study are highly correlated, but there is still a common component among all of the dimensions to represent most of the variance in teachers' scores. Principal Component Analysis (PCA) was used by several studies to investigate the systematic relationship among the eight dimensions of FFT. Their results show that the first principal component with almost equal weight on each dimension explain over 60% of the total variance using Grade 4-8 teachers in the MET study (Garrett & Steinberg, 2014; Kane, Taylor, Tyler, & Wooten, 2011). The purpose of using PCA is to reduce the dimensions that an instrument has to begin with to see whether there is just one major aspect of teacher quality that each dimension contributes to equally to measure, or there are multiple aspects of teacher quality that the instrument tries to capture with all these dimensions, with each contributing different weights towards the measurement of the construct. In the FFT example above, simply averaging all dimensions to get one single composite score for each lesson is meaningful in that the PCA shows that one major component with approximately equal weights accounts for the majority of the variance in the data.

Following this method, I also conducted PCA on the teacher samples in the MET study core files in order to examine and interpret the components generated for all four instruments. The core files include observation data that were scored using the videos from Year One teachers who were also randomized in Year Two of the study. Using the PCA results, I constructed an algorithm by instrument and by subject to generate composite scores to represent the aspects of

teacher quality measured, depending on the component(s) extracted and the meaning of those components. For a situation in which there is only one component extracted for an instrument, I would use the factor loading of each variable to generate the algorithm to quantify the final observational ratings for each lesson, and then get the final teacher level ratings. For a situation in which there is more than one component extracted, I would calculate the component scores for each lesson, and then compare the component scores at the teacher level across instruments and content. With the MET data, all instruments have multiple components extracted[6]. These components are used as the basis in analysis of differences in teachers' quality as manifested in a variety of contexts and content.

4.2.1. Framework for Teaching (FFT)

For lessons rated using FFT, I considered mathematics lessons and ELA lessons separately, but the results are highly consistent across both subjects. For PCA on all ELA lessons in the core files, there are two principal components extracted with corresponding eigenvalue greater than one (Kaiser Criterion). The second largest eigenvalue is only slightly larger than one ($\lambda = 1.027$). But with the inclusion of it, the accumulative variance explained exceeds 60% of the total variance in the data, which conforms to an alternative criterion of selecting components. For PCA on mathematics lessons, however, only one component would be kept when considering the Kaiser Criterion to only keep the component with eigenvalue larger than one. This component has approximately equal factor loadings on each variable within the component. The variance explained is in the 50%~60% range. The second largest eigenvalue is just below

[6] For an introduction of PCA and the processes of factor analysis for these four instruments, please refer to Appendix B.

one ($\lambda = 0.982$). In this case, I have decided to force the second component to be extracted so that the overall variance explained is over 60% (in the 60% ~ 70% range).

I further tested the sensitivity of the PCA results using different sub-samples[7] as well as the full sample of the core files data. The results are highly consistent with the initial PCA results of the subject-specifics samples. Table 9 below summarizes the results of PCA on ELA and mathematics lessons respectively.

Table 9: PCA results of FFT

	ELA			MATH		
	Eigenvalue	% of Variance	Cum. Variance	Eigenvalue	% of Variance	Cum. Variance
COMP1	4.492	56.146%	56.146%	4.467	55.839%	55.839%
COMP2	1.027	12.836%	69.982%	0.982	12.273%	68.112%
COMP3	0.501	6.268%	75.250%	0.519	6.483%	74.596%
COMP4	0.472	5.905%	81.155%	0.498	6.230%	80.826%
COMP5	0.440	5.501%	86.656%	0.465	5.813%	86.639%
COMP6	0.389	4.861%	91.517%	0.419	5.240%	91.879%
COMP7	0.343	4.285%	95.802%	0.339	4.237%	96.116%
COMP8	0.336	4.198%	100.000%	0.311	3.884%	100.000%

Principal Components (Eigenvectors and Rotated Eigenvectors)

	ELA				Math			
	Initial		Rotated		Initial		Rotated	
FFT	COMP1	COMP2	COMP1	COMP2	COMP1	COMP2	COMP1	COMP2
2a	0.756	0.367	0.346	0.766	0.749	0.342	0.359	0.762
2b	0.806	-0.134	0.706	0.412	0.797	-0.146	0.698	0.410
2c	0.713	0.434	0.271	0.789	0.723	0.423	0.271	0.792
2d	0.696	0.550	0.183	0.868	0.717	0.534	0.194	0.873
3a	0.754	-0.095	0.640	0.409	0.742	-0.084	0.617	0.422
3b	0.738	-0.398	0.822	0.166	0.711	-0.393	0.796	0.167
3c	0.785	-0.291	0.790	0.279	0.785	-0.297	0.789	0.288
3d	0.740	-0.362	0.801	0.194	0.728	-0.359	0.786	0.204

Note: 2a: Creating an environment of respect and rapport; 2b: Establishing a culture for learning; 2c: Managing classroom procedures: 2d: Managing student behaviors; 3a:

[7] The sub-samples include: 1) sample of all 2010 data in the core files, 2) sample of all 2011 data in the core files, 3) sample of 2010 mathematics lessons data in the core files, 4) sample of 2011 mathematics lessons data in the core files, 5) sample of 2010 ELA lessons in the core files, 6) sample of 2011 ELA lessons data in the core files.

Communicating with students; 3b: Using questioning and discussion techniques; 3c: Engaging students in learning; 3d: Using assessment in instruction.

I used Varimax with Kaiser Normalization for the orthogonal rotation of the components[8]. After rotation, the first component has higher loadings on *2b Establishing a culture for learning*; *3b Using questioning and discussion techniques*; *3c Engaging students in learning; and 3d Using assessment in instruction*. The second component has high loadings on *2a Creating and environment of respect and rapport*; *3c Managing classroom procedures*; *2d Managing student behaviors*. The first component focuses more on ***Instruction***, while the second component focuses more on ***Management***. Specifically, *3a Communicating with students* has relatively high loadings on both components, which can be interpreted in terms of the types and content of communication with students with respect to different components (the instructional communication vs. the managerial communication with students). After rotation, the first component roughly explains about 30%~40% of the total variance rather than more than 50%, while the second component explains about 20%~30% of the total variance. The total accumulative variance explained by the two components remains the same after rotation of the rotation of the original two components.

The formulae to calculate the component scores of the first and the second principal component for the ELA lessons are:

$$ELA_{Instruction} = 0.346 \cdot 2a + 0.706 \cdot 2b + 0.271 \cdot 2c + 0.183 \cdot 2d + 0.640 \cdot 3a + 0.822 \cdot 3b + 0.790 \cdot 3c + 0.801 \cdot 3d$$

[8] Other orthogonal rotation methods—Equimax and Quartimax—result in similar loadings and component interpretations. Varimax improves the factor pattern equally good or better than the other rotation methods performed.

$$ELA_{Management}$$

$$= 0.766 \cdot 2a + 0.412 \cdot 2b + 0.789 \cdot 2c + 0.868 \cdot 2d + 0.409 \cdot 3a + 0.166 \cdot 3b + 0.279 \cdot 3c + 0.194 \cdot 3d$$

In the above formulae, *2a, 2b, 2c, 2d, 3a, 3b, 3c*, and *3d* are the ratings of the corresponding dimensions in FFT that teachers received for each ELA lesson.

The formulae to calculate the component scores of the first and the second principal component for the mathematics lessons are:

$$Math_{Instruction}$$

$$= 0.359 \cdot 2a + 0.698 \cdot 2b + 0.271 \cdot 2c + 0.194 \cdot 2d + 0.617 \cdot 3a + 0.796 \cdot 3b + 0.789 \cdot 3c + 0.786 \cdot 3d$$

$$Math_{Management}$$

$$= 0.762 \cdot 2a + 0.410 \cdot 2b + 0.792 \cdot 2c + 0.873 \cdot 2d + 0.422 \cdot 3a + 0.167 \cdot 3b + 0.288 \cdot 3c + 0.204 \cdot 3d$$

In the above formulae, *2a, 2b, 2c, 2d, 3a, 3b, 3c*, and *3d* are the ratings of the corresponding dimensions in FFT that teachers received for each mathematics lesson.

4.2.2. Classroom Assessment Scoring System (CLASS)

Similar to the PCA of FFT, I considered mathematics lessons and ELA lessons in the core files separately to conduct PCA for CLASS, and tested the sensitivity of the results using different sub-samples[9] as well as the full sample of the core files data. The PCA results are highly consistent across different iterations. In all iterations, two principal components are

[9] The sub-samples include: 1) sample of all 2010 data in the core files, 2) sample of all 2011 data in the core files, 3) sample of 2010 mathematics lessons data in the core files, 4) sample of 2011 mathematics lessons data in the core files, 5) sample of 2010 ELA lessons in the core files, 6) sample of 2011 ELA lessons data in the core files, 7) lessons scored with the elementary version, and 8) lessons scored with the secondary version.

extracted, regardless of the criteria used to decide on the number of components. The first component roughly explains about 40% of the total variance, while the two components combined explain over 60% of the total variance. Table 10 below summarizes the results of PCA on ELA and mathematics lessons respectively.

Table 10: PCA results of CLASS

	ELA			MATH		
	Eigenvalue	% of Variance	Cum. Variance	Eigenvalue	% of Variance	Cum. Variance
COMP1	5.876	48.964%	48.964%	5.721	47.678%	47.678%
COMP2	1.693	14.105%	63.069%	1.709	14.245%	61.924%
COMP3	0.754	6.287%	74.551%	0.761	6.343%	68.267%
COMP4	0.624	5.196%	74.551%	0.628	5.230%	73.498%
COMP5	0.531	4.427%	78.978%	0.582	4.848%	78.346%
COMP6	0.474	3.947%	82.925%	0.488	4.066%	82.411%
COMP7	0.447	3.729%	86.654%	0.464	3.864%	82.411%
COMP8	0.384	3.201%	89.856%	0.397	3.311%	85.586%
COMP9	0.365	3.039%	92.895%	0.365	3.039%	92.625%
COMP10	0.319	2.655%	95.550%	0.316	2.637%	95.262%
COMP11	0.304	2.533%	98.083%	0.308	2.566%	97.828%
COMP12	0.230	1.917%	100.000%	0.261	2.172%	100.000%

Principal Components (Eigenvectors and Rotated Eigenvectors)

	ELA				Math			
	Initial		Rotated		Initial		Rotated	
CLASS	Comp1	Comp2	Comp1	Comp2	Comp1	Comp2	Comp1	Comp2
PC	0.729	-0.011	0.657	0.317	0.737	0.019	0.707	0.207
NC	-0.397	0.653	-0.086	-0.760	-0.430	0.609	-0.224	-0.711
TS	0.760	0.073	0.720	0.253	0.753	0.133	0.730	0.190
RSP	0.721	0.312	0.786	0.020	0.688	0.349	0.761	-0.122
BM	0.489	-0.740	0.132	0.877	0.499	-0.733	0.252	0.850
P	0.517	-0.610	0.212	0.771	0.524	-0.624	0.209	0.754
ILF	0.791	0.059	0.742	0.280	0.780	0.038	0.754	0.202
CU	0.770	0.168	0.769	0.171	0.753	0.133	0.758	0.133
APS	0.711	0.299	0.771	0.028	0.685	0.344	0.755	-0.109
QF	0.822	0.203	0.831	0.162	0.802	0.213	0.829	0.042
ID	0.807	0.253	0.839	0.110	0.778	0.308	0.835	-0.056
SE	0.740	-0.362	0.801	0.194	0.728	-0.359	0.786	0.204

Note: PC: Positive Climate; NC: Negative Climate; TS: Teacher Sensitivity; RSP: Regard for Student Perspectives; BM: Behavior Management; P: Productivity; ILF: Instructional Learning Format; CU: Content Understanding; APS: Analysis and Problem Solving; QF: Quality of Feedback; ID: Instructional Dialogue; SE: Student Engagement.

In order to interpret the components meaningfully, Varimax with Kaiser Normalization was used for component rotation in the ELA case, and Quartimax with Kaiser Normalization was used for component rotation in the mathematics case[10]. After rotation, the first component has high loadings on *Teacher Sensitivity, Regard for Student Perspectives, Instructional Learning Format, Content Understanding, Analysis and Problem Solving, Quality of Feedback, Instructional Dialogue,* and *Student Engagement*; it also has relatively high loadings on *Positive Climate*. Besides *Student Engagement*, which by itself is an independent domain in the original CLASS framework, all of the above mentioned elements are under the domains of *Instructional Support* and *Emotional Support*. Accordingly, the first component is related to efforts to support students' engagement in learning—***Support.*** The second component has high loadings on *Negative Climate* (negative direction), *Behavior Management*, and *Productivity*, which are mainly under the domain of *Classroom Organization* in the original CLASS framework. Accordingly, the second component focuses on the management and organization of the classroom processes—***Organization***. After rotation, the first component explains about 40% of the total variance, which is similar to the variance explained by the first component before rotation. The total variance explained by the two rotated components remains the same by nature of the rotation.

The formulae to calculate the component scores of the first and the second principal component for the ELA lessons are:

---

[10] Different orthogonal rotation methods—Varimax, Equimax, and Quartimax—result in similar loadings and component interpretation. Varimax, however, improves the factor patter equally good or better than the other rotation methods performed for ELA lessons, while Quartimax improves the factor pattern the best for mathematics lessons.

$$ELA_{Support} = 0.657 \cdot PC - 0.086 \cdot NC + 0.720 \cdot TS + 0.786 \cdot RSP + 0.132 \cdot BM + 0.212 \cdot P + 0.742 \cdot ILF + 0.769 \cdot CU + 0.771 \cdot APS + 0.831 \cdot QF + 0.839 \cdot ID + 0.801 \cdot SE$$

$$ELA_{Organization} = 0.371 \cdot PC - 0.760 \cdot NC + 0.253 \cdot TS + 0.020 \cdot RSP + 0.877 \cdot BM + 0.771 \cdot P + 0.280 \cdot ILF + 0.171 \cdot CU + 0.028 \cdot APS + 0.162 \cdot QF + 0.110 \cdot ID + 0.194 \cdot SE$$

In the above formulae, *PC, NC, TS, RSP, BM, P, ILF, CU, APS, QF, ID*, and *SE* are the ratings of the corresponding dimensions in CLASS that teachers received for each ELA lesson.

The formulae to calculate the component scores of the first and the second principal component for the mathematics lessons are:

$$Math_{Support} = 0.707 \cdot PC - 0.224 \cdot NC + 0.730 \cdot TS + 0.761 \cdot RSP + 0.252 \cdot BM + 0.309 \cdot P + 0.754 \cdot ILF + 0.758 \cdot CU + 0.755 \cdot APS + 0.829 \cdot QF + 0.835 \cdot ID + 0.656 \cdot SE$$

$$Math_{Organization} = 0.207 \cdot PC - 0.711 \cdot NC + 0.190 \cdot TS - 0.122 \cdot RSP + 0.850 \cdot BM + 0.754 \cdot P + 0.202 \cdot ILF + 0.133 \cdot CU - 0.109 \cdot APS + 0.042 \cdot QF - 0.056 \cdot ID + 0.353 \cdot SE$$

In the above formulae, *PC, NC, TS, RSP, BM, P, ILF, CU, APS, QF, ID*, and *SE* are the ratings of the corresponding dimensions in CLASS that teachers received for each mathematics lesson.

4.2.3. Mathematical Quality of Instruction (MQI)

MQI is a subject-specific instrument that is only used to score mathematics lessons. The dimensions are: *Error & Imprecision (EI), Classroom Work Connected to Mathematics (CWCM), Explicitness & Thoroughness (ET), Student Participation in Meaning Making & Reasoning (SPMMR), Richness (R),* and *Working with Students & Mathematics (WSM)*. Moreover, there is a holistic *Lesson Based Guess at Mathematical Knowledge for Teaching* score and an overall *Mathematical Quality of Instruction* score for the lesson. For PCA analysis of the mathematics lesson score by MQI, however, not all of the ratings in these dimensions are included. The dimension ***CWCM*** is a binary variable indicating whether each 7.5-minute segment of the lesson is mainly mathematical or not, which is on a different scale than the other dimensions (3-point scale to indicate the quality). Additionally, there are many missing cases for both the dimensions of ***ET*** and ***CWCM***, at both the holistic level as well as the segment level[11]. Additionally, ***ET*** is not included in the most current version of MQI rubrics by its developers. Hence, it is reasonable to exclude this dimension in this analysis not only because of the missing values in the core files, but also because of the lack of practical importance of this dimension in teacher evaluation using MQI nowadays. In conclusion, the data used for PCA are those dimensions and their scores at the segment levels with complete data, which include ***SPMMR, WSM, R*** and ***EI***.

In the PCA with the full sample of the core files, the second largest eigenvalue is slightly below one ($\lambda = 0.993$). In this case, since the first principal component only explains approximately 40% of the total variance in the data, I have decided to force the second

[11] In the core files, if the lesson got rated in the ET dimension, it does not get rated in the CWCM dimension. The total missing cases at the segment level are 21,752 for ET, and 5,163 for CWCM, out of a total of 26,664 cases.

component to be extracted so that the overall variance explained is over 60% (in the 60%~70% range). I further tested the sensitivity of the PCA results of different sub-samples[12] within the core files, and the results are highly consistent with the initial PCA results of the full sample. Table 11 below summarizes the PCA results and the components extracted from the mathematics lessons.

Table 11: PCA results of MQI

	INITIAL			ROTATED		
	Eigenvalue	% of Variance	Cum. Variance	Eigenvalue	% of Variance	Cum. Variance
COMP1	1.560	39.004%	39.004%	1.550	38.751%	38.751%
COMP2	0.993	24.829%	63.833%	1.003	25.083%	63.833%
COMP3	0.783	19.581%	83.414%			
COMP4	0.663	4.198%	10.000%			

PRINCIPAL COMPONENTS (INITIAL AND ROTATED)				
	Initial		Rotated	
MQI	COMP1	COMP2	COMP1	COMP2
EI	-0.167	0.982	0.034	0.996
SPMMR	0.753	0.161	0.768	0.059
R	0.675	0.022	0.671	-0.069
WSM	0.715	0.039	0.713	-0.056

Note: EI—Errors & Imprecisions; SPMMR—Student Participation in Meaning Making & Reasoning; R—Richness; WSM—Working with Students & Mathematics.

Component rotation was performed using Varimax with Kaiser Normalization for consistency in methodology and easiness for interpretation. But the loadings do not change much after rotation. The first component has higher loadings on all dimensions except for the Errors & Imprecision, while the second component is essentially just Errors & Imprecision. Hence, the first component focuses more on working with students and mathematics—***Instruction***, while the second component focuses more on ***Accuracy***.

[12] The sub-sample include: 1) sample of all 2010 data in the core files, and 2) sample of all 2011 data in the core files.

The formulae to calculate the component scores of the first and the second principal component for the mathematics lessons are:

$$Math_{Instruction} = -0.034 \cdot EI + 0.768 \cdot SPMMR + 0.671 \cdot R + 0.713 \cdot WSM$$

$$Math_{Accuracy} == 0.996 \cdot EI + 0.059 \cdot SPMMR - 0.069 \cdot R - 0.056 \cdot WSM$$

In the above formulae, ***EI, SPMMR, R,*** and ***WSM*** are the ratings of the corresponding dimensions in MQI that teachers received for each mathematics lesson.

4.2.4. Protocol for Language Arts Teaching Observation (PLATO)

PLATO is a subject-specific instrument that is only used to score ELA lessons. Three components are extracted based on Kaiser Criteria of keeping the component with corresponding eigenvalue larger than one. The three components explain about 76% of the total variance, with the first two components explaining about 58%. If the numbers of component is decided based on the total variance explained as the decision rule used for the other three instruments (over 60%), there are still three components extracted. Besides conducting PCA on the full sample of the core file, I further tested the sensitivity of the PCA using different sub-samples[13] within the core files. The results are highly consistent with the initial PCA results of the full sample. Table 12 below summarizes the results of PCA and the components extracted from the ELA lessons.

[13] The sub-samples include: 1) sample of all 2010 data in the core files, and 2) sample of all 2011 data in the core files.

Table 12: PCA results of PLATO

	INITIAL			ROTATED		
	Eigenvalue	% of Variance	Cum. Variance	Eigenvalue	% of Variance	Cum. Variance
COMP1	2.259	37.551%	37.551%	1.596	26.602%	26.602%
COMP2	1.225	20.411%	58.062%	1.003	24.852%	51.454%
COMP3	1.089	18.154%	76.216%	1.486	24.752%	76.216%
COMP4	0.506	8.437%	83.653%			
COMP5	0.499	8.313%	92.966%			
COMP6	0.663	4.198%	10.000%			

PRINCIPAL COMPONENTS (INITIAL AND ROTATED)						
	Initial			Rotated		
PLATO	COMP1	COMP2	COMP3	COMP1	COMP2	COMP3
IC	0.693	-0.075	0.547	0.872	0.113	0.113
CD	0.693	-0.089	-0.543	0.870	0.103	0.123
M	0.515	0.638	0.302	0.046	0.871	0.055
SUI	0.594	0.577	0.216	0.168	0.845	0.086
TM	0.639	-0.454	0.334	0.220	0.098	0.818
BM	0.521	-0.514	0.496	0.027	0.047	0.882

Note: IC— Intellectual Challenge; CD—Classroom Discourse; M—Modeling; SUI—Strategy Use and Instruction; TM—Time Management; BM—Behavior Management.

Component rotation is performed using Varimax with Kaiser Normaliation[14] to clarify the factor pattern. The first component has high loadings on *Intellectual Challenge* and *Classroom Discourse*, which focus on ***Access*** to rigorous content. The second component has high loadings on *Modeling and Strategy Use and Instruction*, which focus on teaching ***Practices***. The third component has high loadings on *Time Management* and *Behavior Management*, which focus on classroom ***Management***. After rotation, the variance explain by each component is approximately equal to each other.

The formulae to calculate the component scores the three components for the ELA lessons are:

[14] Other orthogonal rotation methods—Equimax and Quartimax—were also used to clarify the factor pattern, but both methods result in essentially identical loadings on variables as in Varimax.

$$ELA_{Access} = 0.872 \cdot IC + 0.870 \cdot CD + 0.046 \cdot M + 0.168 \cdot SUI + 0.220 \cdot TM + 0.027 \cdot BM$$

$$ELA_{Practices} = 0.113 \cdot IC + 0.103 \cdot CD + 0.871 \cdot M + 0.835 \cdot SUI + 0.098 \cdot TM + 0.047 \cdot BM$$

$$ELA_{Management} = 0.113 \cdot IC + 0.123 \cdot CD + 0.055 \cdot M + 0.086 \cdot SUI + 0.818 \cdot TM + 0.882 \cdot BM$$

In the above formulae, ***IC, CD, M, SUI, TM***, and ***BM*** are the ratings of the corresponding dimensions in PLATO that teachers received for each ELA lesson.

4.2.5. Summary of PCA Algorithm to Generate Composite Scores

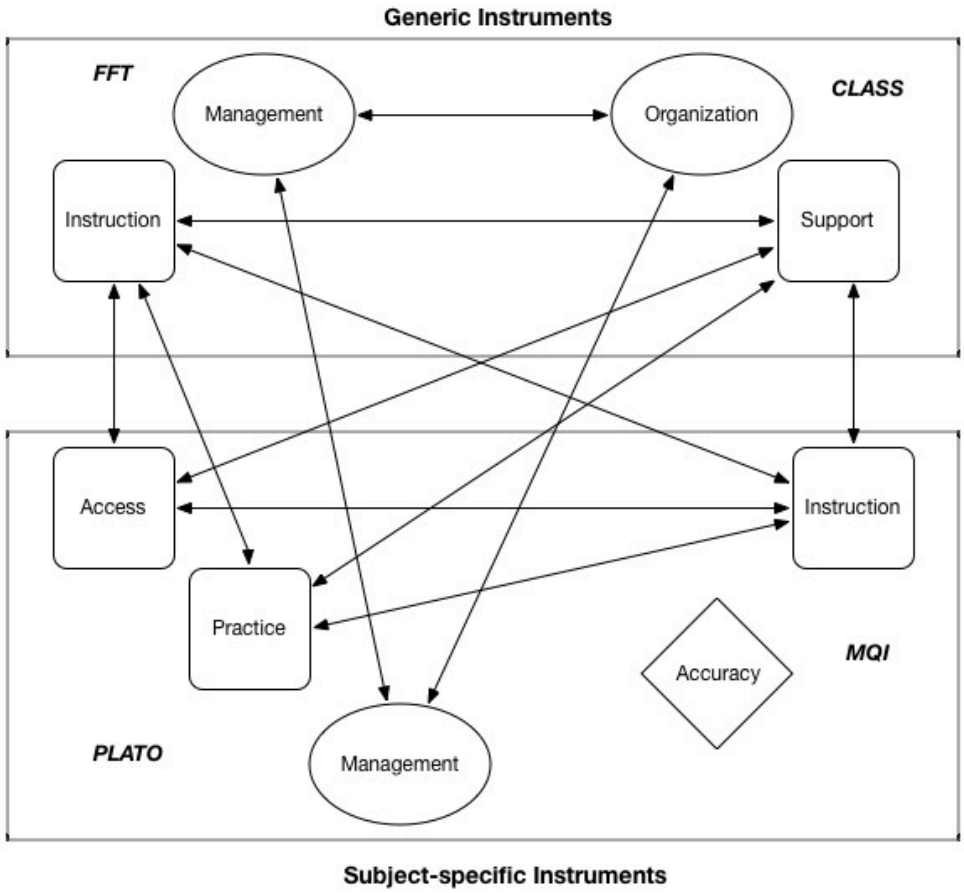
Using the formulae by subject and instrument in this chapter, each lesson receives two or three component scores as the representation of certain aspects of teacher quality measured by a particular instrument. Each of the component score is a composite score of all dimensions with different weight in the instrument. There are two types of aggregation in order to get ratings of teachers at each dimension for each instrument before calculating the PCA component scores with the above formulae. First, for those instruments that rate lessons at the segment level (CLASS and MQI), the segment ratings of each dimension were aggregated within the lessons first before entering the formulae as variables. Second, for lessons rated by two raters, the ratings from each rater were also aggregated at each dimension within the instrument before using them for component score calculation. To be more specific, for generalist teachers, all of their ELA lessons' component scores are aggregated to compare with the components scores aggregated from their mathematics lessons. Similarly, for mathematics teachers, all of the subject areas' component scores, say, scores attained from Algebra & Algebraic Thinking lessons, are aggregated to compare with the aggregated scores of all of their Numbers & Operations lessons.

In summary, the multiple component scores of each instrument generated from the PCA algorithm are first calculated at the lesson level, then they are aggregated at the teacher level based on the different grouping of teachers and subsets of teachers' lessons to get the final ratings for comparisons.

4.3. Discussion and Component Mapping Across Instruments

Based on the PCA results, each instrument measures more than one aspect of teacher quality, and different instruments have different focuses and weights on what is important in evaluating teachers' practices. FFT has eight different dimensions to begin with, but it essentially measures two aspects of teaching: ***Instruction*** and classroom ***Management***. CLASS has twelve dimensions to begin with, but it essentially measures two aspects of teaching: Instructional and emotional ***Support*** and classroom ***Organization***. MQI-Lite has four dimensions at the segment levels to begin with, but it essentially measures two aspects of mathematics teaching: ***Instruction*** and ***Accuracy*** of the content. PLATO has six dimensions to begin with, but it essentially measures three aspects of ELA teaching: ***Access*** to rich content, teaching ***Practices***, and classroom ***Management***. In different subjects, the weights on the original set of dimensions vary, but not drastically, in the compositions of the components. Some of the components extracted are similar in conceptualization across different instruments, while the ***Accuracy*** component from MQI is a stand-alone sub-construct by itself and does not relate to other components conceptually. The following conceptual mapping (Figure 2) shows the relationship among components across instruments. In the figure, the same shapes are connected to represent similar sub-constructs of teacher quality measured across instruments. Those pairs are used for comparison in the analyses.

Figure 2: Conceptual mapping of related components across instruments



4.4. Descriptive Statistics of Composite Scores Generated by PCA and Simple Average

The descriptive statistics in this section represent an overview of the distributions of the generalist and mathematics teachers' observational scores attained from the PCA and simple average algorithms respectively at the aggregated level across different instruments and content areas. For the distribution of all four instruments by year, subject, and composite score aggregation algorithm, please see Appendix B. In general, the distributions of component scores from FFT and CLASS approximate normal distribution, suggesting parametric tests are appropriate for mean comparison. But for MQI, both components' score distributions are very

right-skewed with the majority of the scores clustering around the lower end, suggesting non-parametric tests are appropriate for means comparison.

The descriptive statistics of component scores and the simple averages are as follows, with generalist teachers in Table 13, and mathematics teachers in Table 14.

Table 13: Mean component scores of generalist teachers' aggregated ELA and mathematics lessons

		Year 1			Year 2		
Instrument	Principal Component	N	Mean	SD	N	Mean	SD
FFT	ELA_Instruction	440	11.697	1.159	313	11.693	1.252
	ELA_Management	440	10.642	0.908	313	10.576	0.988
	ELA Average	440	2.641	0.238	313	2.633	0.259
	Math_Instruction	440	11.392	1.208	313	11.371	1.286
	Math_Management	440	10.539	1.056	313	10.448	1.090
	Math Average	440	2.595	0.262	313	2.582	0.274
CLASS	ELA_Support	440	29.915	3.286	313	29.852	2.914
	ELA_Organization	440	15.800	1.484	313	15.662	1.286
	ELA Average	440	4.599	0.399	313	4.583	0.356
	Math_Support	440	29.907	3.445	313	30.041	3.158
	Math_Organization	440	12.834	1.322	313	12.658	1.194
	Math Average	440	4.539	0.424	313	4.540	0.390
MQI	Math_Instruction	430	2.566	0.335	310	2.734	0.317
	Math_Accuracy	430	1.145	0.222	310	1.150	0.198
	Math Average	430	1.572	0.137	310	1.628	0.128
PLATO	ELA_Access	430	5.504	0.678	310	5.642	0.624
	ELA_Practices	430	4.142	0.785	310	4.067	0.750
	ELA_Management	430	7.213	0.518	310	7.332	0.477
	ELA Average	430	1.750	0.261	310	1.708	0.237

Table 14: Mean component scores of mathematics teachers' aggregated lessons

		Year 1			Year 2		
Instrument	Principal Component	N	Mean	SD	N	Mean	SD
FFT	Math_Instruction	978	10.900	2.060	772	10.895	2.143
	Math_Management	978	10.079	1.814	772	10.039	1.858
	Math_Average	978	2.484	0.332	772	2.478	0.327
CLASS	Math_Support	978	27.950	4.247	772	28.138	4.011
	Math_Organization	978	12.284	1.721	772	12.193	1.537
	Math_Average	978	4.298	0.531	772	4.314	0.499
MQI	Math_Instruction	971	2.541	0.529	770	2.671	0.492
	Math_Accuracy	971	1.124	0.345	770	1.127	0.342
	Math_Average	971	1.568	0.129	770	1.613	0.120

4.5. The Distributions of Observational Ratings by Subject Areas for Mathematics Teachers

This section focuses on the breakdown of mathematics lessons by subject areas and the respective distribution of the components scores and the simple average scores. The distribution of the observational scores at the lesson level provides a holistic picture on the scores and the variation that lessons in a particular area of mathematics comparing with lessons in other areas. Each component and the simple average within the instrument are at a different scale due to the construction methods described in previous sections of this chapter. Hence it is only meaningful to compare them within subject areas at each comparison level. I put them side-by-side to give readers an idea of the relative scale among these three types of measures generated by the same instrument as indicators of teacher quality. Noted that in later chapters, the smaller the scale, the more clustered the scores, and hence it is more difficult to detect difference scores of the absolute values.

Figure 3: FFT raw component scores and simple average composite scores across subject areas by year

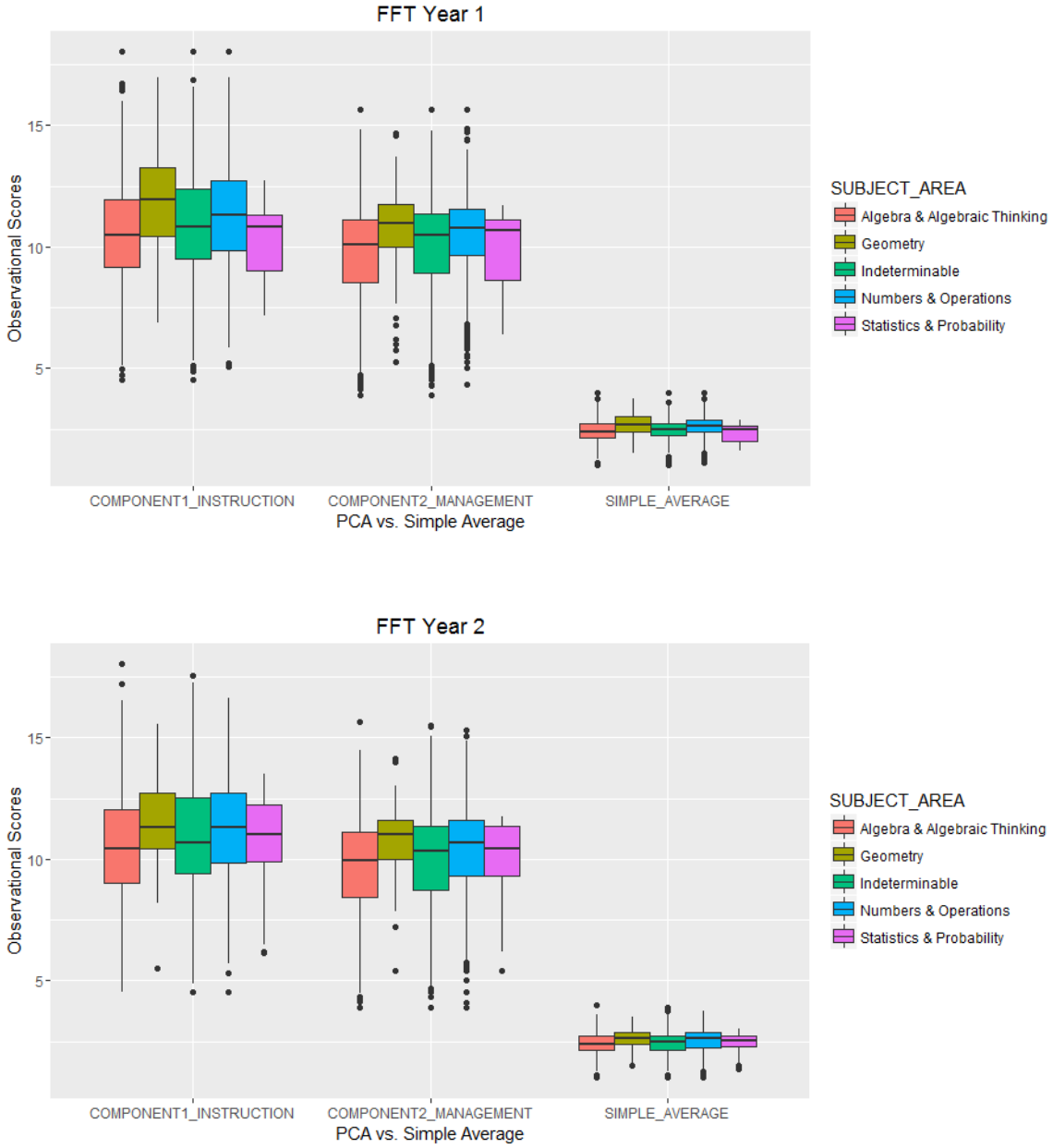


Figure 4: CLASS raw component scores and simple average composite scores across subject areas by year

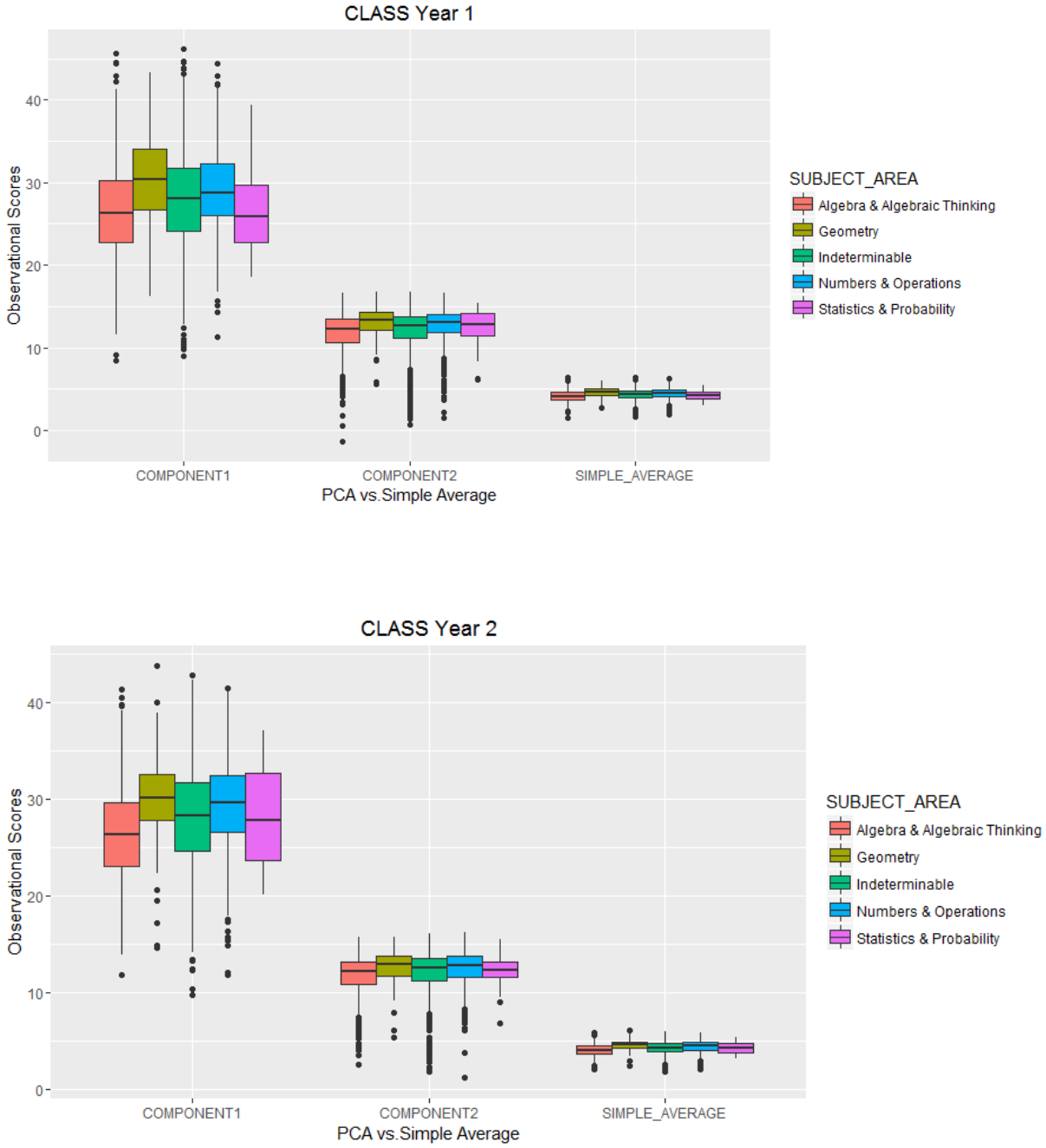
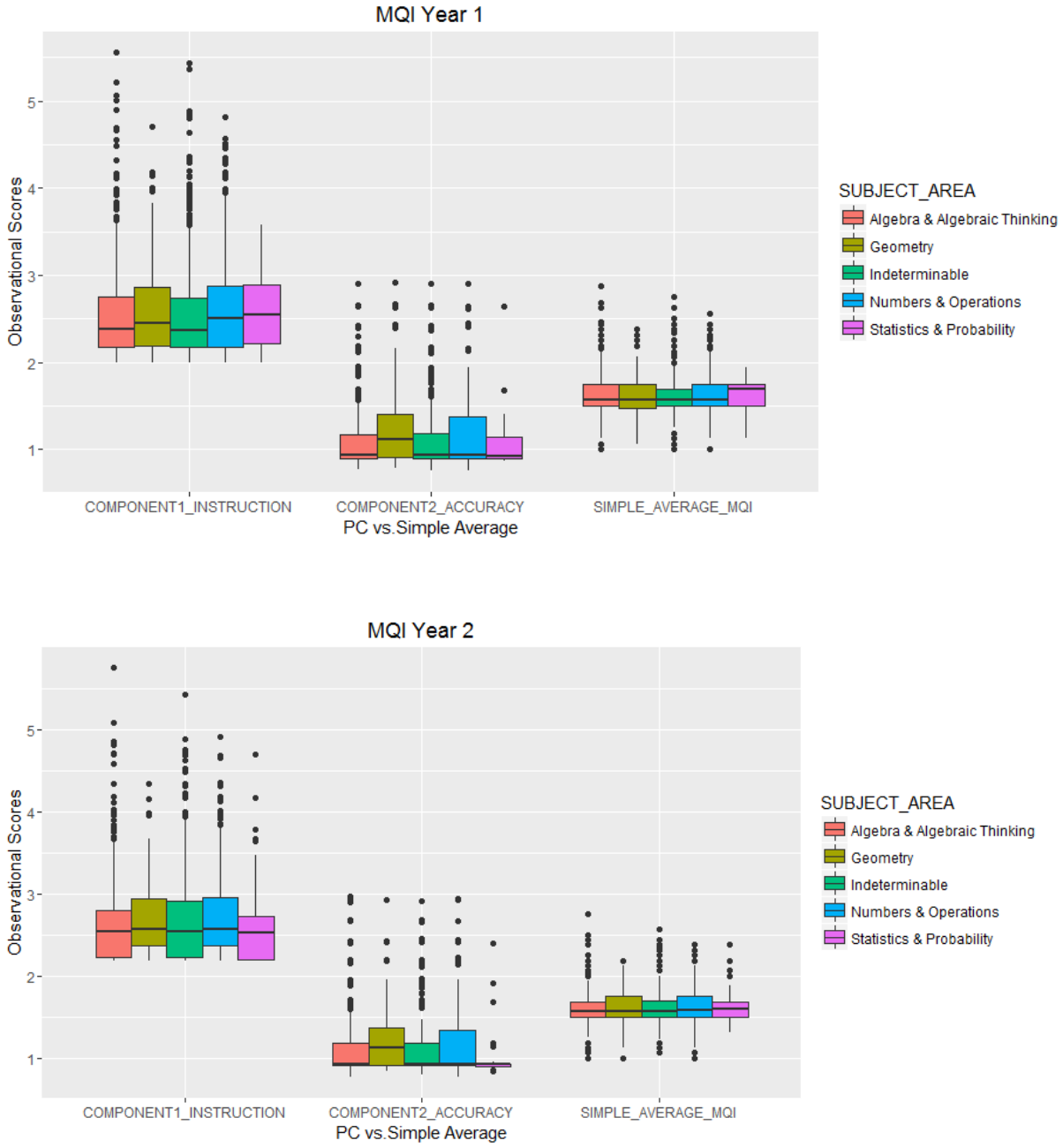


Figure 5: MQI raw component scores and simple average composite scores across subject areas by year



The "indeterminable" category consists of lessons labeled as Random Topic in the data, which encompasses a conglomerate of lessons from all subject areas except for Statistics & Probability (as lessons of this subject area were identified from the Random Topic lessons in the

first place, as described in Chapter 3). As seen in the figures above, in general, there are some variations in the observational scores for mathematics lessons in different subject areas. But at the same time, lessons of different content vary greatly in amount as well. Algebra & Algebraic Thinking and Numbers & Operations lessons make up the majority of the total mathematics lessons, while Geometry and Statistics & Probability lessons are much fewer in numbers, especially the latter subject area. Excluding the lessons whose subject areas cannot be determined, the means among the four unequal groups are significant different (p-value < 0.05)[15] from each other, for various component scores as well as the simple average scores attained from each instrument.

In the next three chapter, the analyses and results to answer the research questions zoom into the lessons taught by the same teacher. I compare those teachers' matched observational ratings at the teacher level across instruments and content in order to understand whether teachers' observational scores attained in different content and contexts are unstable in a way that matters to teacher evaluation.

[15] Both one-way ANOVA to test group means and Kruskal-Wallis Non-parametric to test group distribution difference were used, and all of the tests show statistical significance results among four subject areas.

CHAPTER 5

DIFFERENCES IN GENERALIST TEACHERS' OBSERVATIONAL RATINGS ACROSS SUBJECTS

5.1. Introduction

This chapter addresses the first research question: For the generalist teachers who teach both ELA and mathematics in elementary grades, to what extent are their observational scores different across subjects as measured by various protocols in the MET data? In teacher evaluation, if there were evidence that supported the influence of subjects on generalist teachers' observational ratings, the implicitly held assumption that there is a unifying teacher quality or qualities across subjects for the same teacher should be challenged. If generalist teachers do not get rated similarly when they are teaching different subjects, neglecting the subject matter of the observations may result in inappropriate and unfair evaluation results. In other words, the validity of the inference a stakeholder makes with regard to teachers is at stake if he or she does not differentiate the subject of the observations from which the scores are attained.

The comparisons I make[16] to investigate the answer to this question center around three contextual factors that are relevant in teacher evaluation systems. First, I consider whether a certain instrument used to get the observational ratings matters. In this study, the scores from both the generic instruments and subject-specific instruments are compared and contrasted with oneself and with each other to examine consistency between various observational measures. The second factor concerns the different ways to generate composite scores for the use of evaluation. In this study, I explore two different methods to aggregate scores: PCA and simple

[16] For the list of comparison pairs, please refer to Appendix D.

average (whose methods and rationale are elaborated in Chapter 4), and see whether the results are different when dissimilar approaches are taken.

Lastly, I examine whether the various models to use the observational ratings in order to get the evaluation results can be influential as well. In particular, I used two different perspectives to compare teachers' observational ratings based on two major forms of incentive structure used in the education sector to evaluate teachers: fixed performance contract and rank-order tournament (OECD, 2009). Under the fixed performance contract framework, teachers are assessed in teacher evaluation based on their absolute performance. Cut-off points for each level of performance/quality are created to put teachers into categories without restraining the number of teachers at each level. In this study, the absolute performance is the raw observational scores teachers get in each subject/subject area within mathematics from a particular instrument. Under the rank-order tournament framework, teachers are assessed in teacher evaluation based on their relative performance. In this study, the relative performance is the rankings of the same group of teachers' observational ratings across different contexts as discussed in the first and second considerations. The results in the next sections are presented around these two different frameworks of teacher evaluation systems in order to examine whether the different use of the observational scores lead to dissimilar evaluation results in each scenario.

5.2. The Influence of Subjects on Generalist Teachers' Observational Raw Scores

Analyses using paired-sample t-tests of the MET data show that subject of the lessons has great influence on generalist teachers' observational scores. Of the twelve comparison pairs involving the two generic instruments, teachers' ratings based on observing their ELA lessons are higher than those based on mathematics lessons for all but two combinations, and the

difference is statistically significant in nine comparisons (see Table 15 for the t-test results and the effect sizes of the mean differences).

Table 15: P-values and effect sizes for generic instruments' raw scores comparison

Instrument/ Content	Composite Score Generation Method	Comparison Level	Year One N=440 p-value (effect size)	Year Two N=313 p-value (effect size)
FFT math vs. FFT ELA	PCA	Instruction	0.000*** (0.241)	0.000*** (0.242)
		Management	0.043* (0.106)	0.023* (0.129)
	Simple average	Overall	0.000*** (0.181)	0.001*** (0.187)
CLASS math vs. CLASS ELA	PCA	Support	0.962	0.300
		Organization	0.000*** (2.373)	0.000*** (2.524)
	Simple average	Overall	0.001*** (0.161)	0.051

Note: *** means that the statistics is significant at the 0.001 level (2-tailed). ** means that difference is significant at the 0.01 level (2-tailed). * means that the difference is significant at the 0.05 level (2-tailed). If not significant, only p-value is provided but not the effect size.

As discussed in Lipsey et al. (2012), any effect sizes larger than 0.1 can be considered substantive in comparisons to other studies with broad measures (such as standardized tests) in the domain of education. Accordingly, generalist teachers' two observational measures in distinct subjects are practically and significantly different in 9 out of the 12 cases, with effect sizes ranging from 0.11 to 2.5. It is also worth noting that in the managerial aspect of teacher quality measured by CLASS feature particularly large differences, with the ELA scores of

generalist teachers more than two standard deviations higher than their mathematics ones[17]. Overall, these comparison results support the influence of the subject matter in deciding generalist teachers' observational ratings, which ultimately affect the teachers' evaluation results under the fixed performance framework.

Given the discrepancy in generalist teachers' observational measures across subjects, one might ponder on the variation of the differences under school contexts. Since previous research found that teachers' observational scores on various instruments uniformly tend to decrease as grade level increases (Mihaly & McCaffrey, 2014), are the difference detected in this study larger in higher grade levels than lower ones within elementary levels, or vice versa? Further investigations into the statistical differences indicate that the existence of the discrepancy and the extent of the difference between ELA and mathematics' observational scores does not depend on grade levels and is prevalent for generalist teachers in all Grade 4 to 6; in other words, in the nine cases where generalist teachers' ELA ratings are significantly higher than their mathematics ones, the discrepancy cannot be explained by the factor of grade level[18].

In conclusion, even though the generic instruments are designed to be content-free and can be used to rate a variety of content, the differentiated ratings from the same teachers suggest

---

[17] The two formulae to generate CLASS ***Organization*** scores for ELA lessons and mathematics lessons partially contribute to the large differences in this comparison. Although both components are called ***Organization***, the composition of each component differs a little across subjects. In particular, the dimension of ***Student Engagement*** has much higher weight in this component for mathematics lessons than for ELA lessons, while the dimension of ***Positive Climate*** has much higher weight in this component for ELA lessons than for mathematics lessons (see., Section 4.2.2). The meaning of the component in each subject changes accordingly, but still, the components are largely loaded on the classroom organizational dimensions. The two formulae for CLASS ***Organization*** component are the two most different ones across all formulae. But overall, the compositions of the component are still similar to each other.

[18] For the ANOVA models examining the main effect and the interaction effect involving subject as well as grade level/district with the generalist teachers, please refer to Appendix I.

an inconsistency between the observational measures to assess a unifying teacher quality or aspects of teacher quality across subjects with distinct natures. Moreover, as taking the simple average of all dimensions of an instrument is a common practice in most of the teacher evaluation systems nowadays, the indiscriminant uses of generalist teachers' scores without paying attention to the subject of the observations may lead to inaccurate evaluation results. If a teacher evaluation system sets up the same cut-off points across subjects in order to place generalist teachers into quality categories, teachers are very likely to be in higher category when evaluators unconsciously observe only their ELA lessons, or more ELA lessons than the mathematics lessons during the evaluation implementation processes. The more complex algorithm of using factor analysis to generate teacher quality measures also conforms to the pattern of higher ELA ratings as compared to the simple average method. The two aggregation methods do not perform differently, as the significance/insignificance in the average score comparisons always accompany by the same result in at least one of the component score comparisons.

5.3. The Influence of Subjects on Generalist Teachers' Observational Rank Scores

As discussed in the previous section, generalist teachers' ELA scores on different instruments are higher than their mathematics ones in almost all cases. One possible scenario for these results is that teachers who score well in ELA lessons would also score high in mathematics lessons, and the extent of the difference is close to a constant across all teachers (say, teachers' ELA ratings are always one point higher than their mathematics ones). If this explanation holds, when the same group of teachers' paired scores in ELA and mathematics are used for rankings within the subject for comparison, the difference will be eliminated and might not have consequences in teachers' evaluation results.

The analyses using generalist teachers' rank observational scores in this section refute such explanation to the results in the previous section. Under the rank-order tournament framework to evaluate teachers, there is a large variability in generalist teachers' observational ranking across subjects in all comparison pairs within instruments and score generation algorithms[19]. The following sections will focus on the quantification of the variability in the comparisons between FFT and CLASS first, and then turn to the comparisons of rank scores results from PLATO and MQI.

5.3.1. Generalist Teachers' Rank Scores from Generic Instruments

For the same set of comparisons regarding generic instruments analyzed under the ranking framework, I examine the differences between teachers' two observational scores by ordering the scores within the subject first, and then dividing teachers into bins with approximately equal shares[20] in order to identify teachers' change (or the lack thereof) in those percentile groups. A transition matrix, which is a tool usually used to examine the inter-temporal reliability between teacher quality measures in one year and the next (Aaronson, Barrow, & Sander, 2007; Ballou, 2005; Pivovarova & Amrein-Beardsley, 2015), is of great value as an

---

[19] When a generalist teacher's ELA lessons were measured by PLATO, and his or her mathematics lessons were assessed by MQI, I cannot compare the two observational scores directly using t-tests because the two measures are at different scales. I used a t-test equivalent non-parametric test, which is essentially comparing the ranks of the two scores.

[20] The number of percentile groups to divide the teachers into is a subjective choice. I thought about dividing teachers into quartile, which coincide with or close to the number of categories that many current teacher evaluation systems put teachers into. But when it comes to making personnel decisions, districts are usually more discrete and conservative, and hence a narrower interval to check the bottom and top percentiles stability and volatility are more relevant. Moreover, for some of the samples in this study, the numbers of teachers are large, and dividing into deciles will be more of a finer grain analysis of the rankings. Based on all these considerations, for sample sizes larger than 100 in this study, by default ten percentile groups (i.e., deciles) are generated. For sample sizes smaller than 100, by default only five percentile groups (i.e., quintiles) are generated.

analytical approach to track the stability of individual teachers' quality when assessed in different situations. For this chapter, the scenario examined is the subject, and Table 16 displays one such analysis with a transition matrix that links decile rankings of teachers' ELA scores with decile rankings of the same teachers' scores in mathematics on FFT. In general, the transition matrix's diagonal elements are of most interest to estimate the stability of teachers' observational measures attained under different contexts. Specifically, the diagonal elements are the calculated percentage of teachers who are ranked and categorized into a certain percentile group under one scenario (i.e., the reference group) and are in the same percentile group under another scenario. If there is a perfect consistency between teacher quality measured in one scenario and the other, in theory the diagonal elements are all 100% and the rest of the cells are 0%. In contrast to this, the most extreme situation is the pure random assignment where teachers assessed in ELA are randomly assigned to a new quality ranking for their measures of mathematics instruction. The result under this scenario is that each cell would contain approximately equal amount of teachers, and hence the percentages would be 10% across the board in the matrix. In reality with measurement errors, none of the extremes are probable. For the diagonal elements, the closer they are to 10%, the less consistent are the two quality rankings. In the same fashion, the closer the diagonal elements are to 100%, the more consistent are the two quality rankings in the comparison. In this illustrative example and those future ones, the diagonal elements are colored in green and the adjacent cells (one percentile group below or above) are colored in blue for better visualization of the distribution.

Table 16: Rankings of Year One generalist teachers' ELA and mathematics scores on FFT Average: Percentage of teachers

Decile Rank in math	Decile Rank in ELA (N = 440)									
	1	2	3	4	5	6	7	8	9	10
1	40.9%	14.9%	7.9%	6.7%	16.7%	4.2%	2.4%	0.0%	2.5%	0.0%
2	22.7%	12.8%	10.5%	20.0%	14.3%	10.4%	4.9%	2.0%	7.5%	0.0%
3	13.6%	14.9%	13.2%	13.3%	14.3%	6.3%	0.0%	11.8%	7.5%	6.8%
4	0.0%	17.0%	10.5%	13.3%	2.4%	10.4%	14.6%	7.8%	7.5%	4.5%
5	9.1%	10.6%	5.3%	13.3%	7.1%	6.3%	14.6%	11.8%	0.0%	11.4%
6	2.3%	6.4%	7.9%	11.1%	11.9%	14.6%	12.2%	21.6%	20.0%	4.5%
7	0.0%	6.4%	18.4%	2.2%	9.5%	14.6%	19.5%	9.8%	12.5%	13.6%
8	9.1%	6.4%	18.4%	6.7%	11.9%	16.7%	9.8%	11.8%	15.0%	9.1%
9	0.0%	6.4%	2.6%	8.9%	7.1%	6.3%	7.3%	9.8%	15.0%	22.7%
10	2.3%	4.3%	5.3%	4.4%	4.8%	10.4%	14.6%	13.7%	12.5%	27.3%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Note: ELA is the reference group.

In the above transition matrix where ELA is the reference group, the first element 40.9% in green means that of the 44 teachers who were ranked in the bottom of the population (Decile 1) based on their ELA instruction, 40.9% of them (18 teachers) remained in the bottom when the same group of teachers were ranked based on their mathematics instruction. Moreover, within the first column, 9.1% of the 44 teachers (four teachers) with the poorest performance in their ELA instruction as measured by FFT, were placed in Decile 8 when one evaluates them based on their mathematics instruction, suggesting that they have higher FFT ratings than 70% of the other generalist teachers within the population. One of them even made it to the Decile 10, which

means that he or she is considered as the top performing mathematics teachers within the population[21].

Overall, only 17.5% of the 440 teachers stay in the same percentile groups as evaluated in different subjects by FFT. The bottom and top deciles feature slightly higher percentages of teachers, suggesting teacher quality measures on FFT at the two ends of the spectrum are more consistent with each other. That is, highly unsatisfactory performance and highly satisfactory performance by teachers are more consistent when the same generalist teachers are teaching different subjects. Of the teachers who change their percentile group in the quality distribution, 22.7% of the teachers move up or down one percentile group, while 59.8% of the teachers move up or down at least two or more percentile groups in the distribution[22]. This is approximately one quality category or more in most actual teacher evaluation systems, as districts usually categorize teachers into four or five groups.

The rest of the rank score comparisons I make[23] regarding different methods to aggregate scores as well as another generic instrument—CLASS— have similar results as the one displayed above. Even though with the data, a chi-square test does not suggest the rankings are merely random assignments, the results emphasize the large inconsistency of the two teacher quality measures across subjects within both generic instruments.

---

[21] Future transition matrices in this study follow the same format and interpretations where the reference group noted is the the baseline for comparison; the percentage is calculated based on the sizes of each decile of the reference group.
[22] The information on the numbers and percentages of teachers moving up or down one to nine percentile group for each comparison pair can be found in Appendix H.
[23] For the exact numbers of teachers in each percentile group and the count of teachers who move up or down 1 to 9 groups in the distribution, please refer to the respective comparison pairs in Appendix G and Appendix H.

5.3.2. Generalist Teachers' Rank Scores from Subject-specific Instruments

Would subject-specific instruments tell a different story about generalist teachers' quality measures in different subjects? Do generalist teachers' quality or qualities assessed by subject-specific instruments at the same level across subject matter? The short answer to these questions is no. In fact, there is a larger inconsistency between generalist teachers' quality measures in ELA versus in mathematics when they are assessed by subject-specific instruments respectively, on both the matching components as well as the simple averages. In this section, I demonstrate the results from comparing generalist teachers' ELA and mathematics lessons as assessed by PLATO and MQI specifically (see Appendix D for the complete list of comparisons). This additional set of comparisons (as well as later analyses involving mathematics teachers' MQI scores) provides empirical evidence on the use of subject-specific instruments in teacher evaluation systems, and add to the argument to see whether there are reasons that the districts should switch to subject-specific instruments at extra cost to evaluate teachers, following the trend in the research community.

Overall, none of the six comparison pairs regarding generalist teachers' rank scores assessed in subject-specific instruments demonstrate any strong correlation. The positive associations between the two measures compared, though significant largely because of the sample size, only have negligible to weak correlation coefficients (from 0.15 to 0.23) as quantified by Spearman Rank Order tests[24]. The low correlations demonstrate the robustness of the inconsistency across years and different ways to aggregate scores for generalist teachers' two measures in question.

---

[24] For the scatterplots and the results of Spearman Rank Order tests in this chapter, please refer to Appendix E.1.

Additionally, the volatility between generalist teachers' two measures attained from the subject-specific instruments are larger than the generic ones, as evidenced by the smaller elements in the transition matrices for each comparison, especially along the diagonal. The transition matrix below (Table 17) is a typical case of such instability and the extent to which the two measures are inconsistent across subjects.

Table 17: Rankings of Year One generalist teachers' ELA scores on PLATO Average and mathematics scores on MQI Average: Percentage of Teachers

Decile Rank in math	Decile Rank in ELA (N = 430)									
	1	2	3	4	5	6	7	8	9	10
1	15.9%	8.9%	5.1%	17.4%	9.1%	11.9%	8.1%	10.6%	0.0%	4.7%
2	9.1%	17.8%	7.7%	15.2%	13.6%	7.1%	13.5%	8.5%	9.3%	14.0%
3	13.6%	4.4%	10.3%	4.3%	9.1%	16.7%	5.4%	6.4%	0.0%	2.3%
4	20.5%	11.1%	12.8%	10.9%	11.4%	9.5%	16.2%	10.6%	11.6%	7.0%
5	6.8%	13.3%	2.6%	2.2%	20.5%	11.9%	10.8%	8.5%	14.0%	7.0%
6	11.4%	2.2%	15.4%	15.2%	9.1%	2.4%	5.4%	14.9%	11.6%	0.0%
7	6.8%	15.6%	12.8%	19.6%	6.8%	14.3%	5.4%	12.8%	4.7%	16.3%
8	2.3%	11.1%	15.4%	4.3%	4.5%	7.1%	18.9%	8.5%	7.0%	7.0%
9	9.1%	6.7%	12.8%	4.3%	11.4%	9.5%	5.4%	6.4%	18.6%	20.9%
10	4.5%	8.9%	5.1%	6.5%	4.5%	9.5%	10.8%	12.8%	23.3%	20.9%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Note: ELA is the reference group.

Overall, only 14.6% of the 430 teachers remain in the same decile as evaluated by different instruments and in different subjects. The percentages in the top and bottom deciles are no better than the rest of the percentages on the diagonal, featuring only 15.9% and 20.9% of the teachers remaining in the same place respectively. These low numbers at the two ends are alarming as these percentile groups are the people on whom policymakers make decisions regarding retention and compensation. Additionally, higher percentage of teachers—67.9%—changes at least two or more percentile groups as the subject and instrument change from ELA to mathematics and from PLATO to MQI as compared to the generic instruments. This is a large

number for attention as it signals that near 70% of the generalist teachers would have teacher evaluation results that are one or more quality category higher or lower when they are rated by subject-specific instruments and ranked separately for these two subjects they teach.

5.4. Chapter Summary

In this chapter, I explore several perspectives that are relevant for teacher evaluation practices to compare generalist teachers' ratings in ELA and mathematics. In three quarters to all of the combinations examined, subject plays an imperative role in deciding generalist teachers' observational results, no matter how the dimensional level scores of an observational instrument aggregated into composite scores, no matter what framework is used to categorize teachers. Moreover, under the rank-order tournament framework, the extent of the differences is higher between the two measures of the same teachers when measured by the subject-specific instruments than by the generic ones across the board. In conclusion, the subject matter of a lesson has a great impact on generalist teachers' observational ratings, hence the evaluation results without considering the role played by subject do not present a reliable and stable estimate of the average teacher quality and qualities across subjects for these teachers. Given that the evaluation results are directly tied to teachers' professional development and course of career, the degree of variability in teachers' quality measures that is associated with subject deserves more attention from various parties, including stakeholders and teacher educators.

CHAPTER 6

DIFFERENCES IN MATHEMATICS TEACHERS' OBSERVATIONAL RATINGS ACROSS INSTRUMENTS

6.1. Introduction

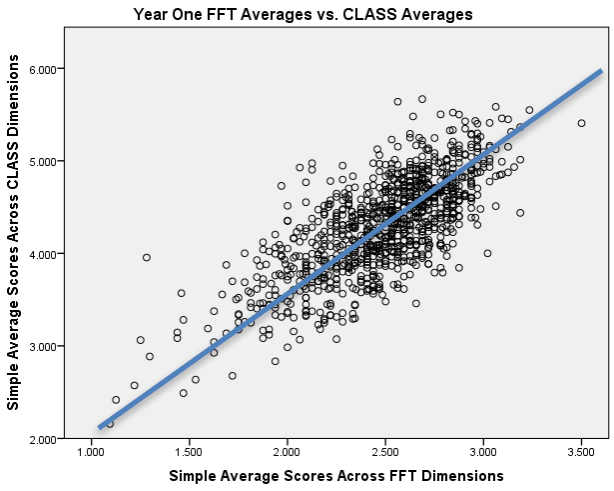
This chapter addresses the second research question: For the mathematics teachers (including those who are generalist teachers in elementary grades and those who teach mathematics specifically in higher grades), to what extent are their observational scores different as measured by various protocols in the MET data? In teacher evaluation, stakeholders are constantly facing the dilemma of selecting the most suitable observational protocol to use among a variety of tools. Generic instruments can be used in classrooms with diverse content, but subject-specific instruments can zoom into content-specific practices, which may be more relevant when one wants to make connections between the teacher performance and the teacher effectiveness measures. But when it comes down to the teacher evaluation results, if dissimilar observational instruments largely agree with each other on determining levels of teachers' quality/qualities, which instrument to use may not as important as other practical considerations, such as cost and personnel training. Hence, the examination of the (in)consistency between different observational instruments in measuring individual teachers' quality or similar aspects of quality is an important step to justify certain choice of a tool in an evaluative context. If there is a lack of consistency in the evaluation results when different instruments are employed, the stakeholders' choice of an observational protocol should be highly aligned with the top priorities in their educational policy agenda. It is also essential to make transparent to teachers and schools how the chosen instrument can help achieve the goal of the district/state more than the others.

In this chapter, I focus on mathematics teachers' quality as measured by three observational instruments with distinct natures in the MET study to compare and contrast their observational ratings. Specifically, I compare the same mathematics teachers' ratings across the two generic instruments, and across one of the generic instruments and the math-specific one. Similar to the previous chapter, I also consider different score aggregation methods when I make the comparisons[25]. The results of the comparisons between the two generic instruments are presented first in the next section, followed by the results of the comparisons between the generic and subject-specific instruments.

6.2. Mathematics Teachers' Observational Ratings Between Generic Instruments

Even though FFT and CLASS come from different theoretical standpoints as observational protocols, the same mathematics teachers' lessons measured by these two instruments are moderately to highly associated with each other. The scatterplot of Year One teacher sample's simple average scores below demonstrates the typical relationship between the FFT measures of mathematics teachers' quality and the CLASS measures of the same construct

Figure 6: Scatterplot of Year One mathematics teachers: FFT Average vs. CLASS Average



[25] For the list of comparisons made for this chapter, please refer to Appendix D.

As seen in the scatterplot, teachers' simple average scores on FFT and CLASS follow a linear pattern and are positively associated. The Spearman rho correlation coefficient ($\rho = 0.694$) also confirms their moderate to high association that can be captured by a monotonic function between the two measures from the same mathematics teacher. The rest of the five comparison pairs between FFT and CLASS also feature a similar linear pattern in their scatterplots and display the same level of associations around 0.7. Such pattern persists across both years' samples to suggest a robustness of the results.

Although the scatterplot demonstrates that there is a relatively high correlation between the two measures, when considering the rank-order tournament framework, a teacher's rank could differ substantially between the two measures, as the slightest nuances in the scores will result in different rankings of teachers when gauging their relative performance against each other. When looking at these mathematics teachers' changes of ranking in the transition matrices[26], we can see that the two measures compared are not always highly consistent throughout the quality ranking distribution, especially for teachers with middle rankings in the population. In order to understand how the medium to high association between the two measures plays out in ranking teachers into categories under the rank-order tournament structure, the transition matrix for the same comparison pair in Figure 7 is presented in Table 18 below:

[26] For all the transition matrices regarding the second research question, please refer to Appendix F.2.

Table 18: Rankings of Year One mathematics teachers' FFT Average scores and CLASS Average scores: Percentage of Teachers

Decile Rank in CLASS	Decile Rank in FFT (N = 978)									
	1	2	3	4	5	6	7	8	9	10
1	58.0%	21.4%	8.3%	6.6%	1.8%	0.0%	1.1%	0.0%	0.0%	0.0%
2	17.0%	27.2%	20.2%	16.0%	7.1%	7.1%	1.1%	1.0%	2.1%	1.0%
3	12.0%	15.5%	13.1%	19.8%	12.4%	10.6%	7.6%	5.8%	1.1%	1.0%
4	8.0%	9.7%	10.7%	16.0%	18.6%	9.4%	10.9%	7.8%	8.4%	0.0%
5	1.0%	9.7%	17.9%	15.1%	11.5%	15.3%	8.7%	12.6%	6.3%	2.1%
6	2.0%	4.9%	10.7%	12.3%	14.2%	14.1%	10.9%	9.7%	14.7%	5.2%
7	1.0%	6.8%	8.3%	7.5%	13.3%	10.6%	15.2%	14.6%	12.6%	9.3%
8	1.0%	0.0%	7.1%	4.7%	9.7%	12.9%	15.2%	19.4%	21.1%	13.4%
9	0.0%	3.9%	2.4%	0.9%	6.2%	15.3%	20.7%	14.6%	16.8%	20.6%
10	0.0%	1.0%	1.2%	0.9%	5.3%	4.7%	8.7%	14.6%	16.8%	47.4%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

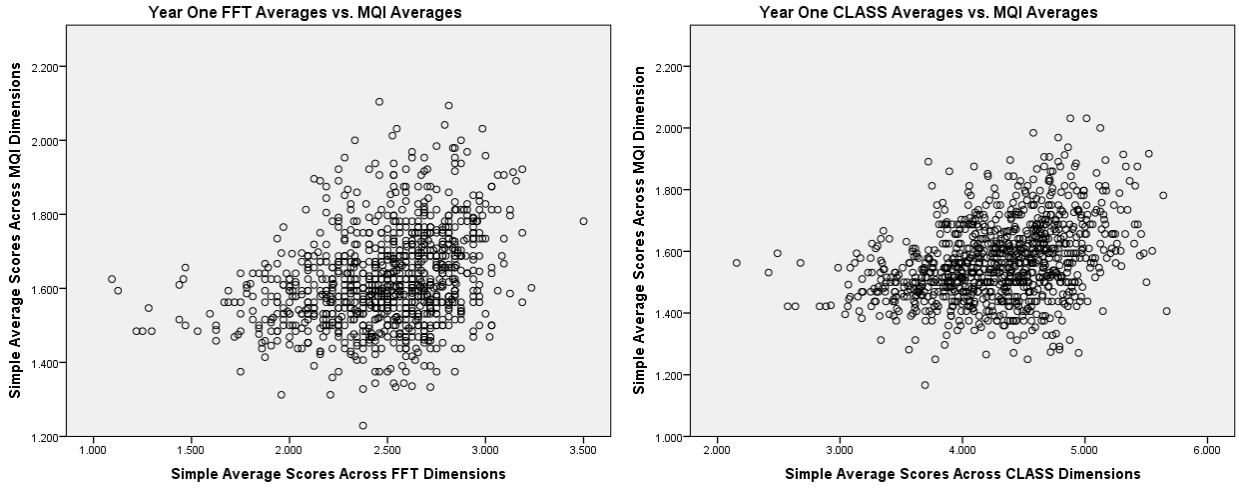
Note: FFT is the reference group.

In this sample analysis, 24% of the 978 teachers remain in the same percentile groups while another 29.3% of them move up or down just one percentile group in the distribution. That is, more than half of the teachers' quality rankings are relatively stable when the measurement changes. The bottom and top percentile groups feature high percentages when considering the diagonal and the adjacent cells together, ranging from 65% to 75% combined. This indicates that the same mathematics teacher's quality measures from the two generic instruments are more consistent with each other at the two ends of the spectrum involving highly unsatisfactory performance and highly satisfactory performance than the middle of the distribution. There are, however, still non-trivial volatility in the middle range of the distribution as evidenced by some low percentages that are not much higher than 10% on the diagonal; and the instability is also evidenced by the fact that (46.7%) of the teachers in the sample move up and down at least two or more percentile groups. This suggests that near half of the teachers' evaluation results would be one or more quality category higher or lower under the rank-order tournament context.

6.3. Mathematics Teachers' Observational Ratings Between Generic and Subject-specific Instruments

In contrast to the results between the two generic instruments, mathematics teachers' quality measures between one of the generic instruments and the subject-specific one are only weakly associated with each other, regardless of what aspect of teacher quality one attends to and how scores are aggregated. One can see this relationship intuitively between the two types of measures from the scatterplots below, in which the simple average scores from each of the generic instruments is plotted against the simple average scores from MQI respectively.

Figure 7: Scatterplots of Year One mathematics teachers: FFT Average or CLASS Average vs. MQI Average



The two comparisons from left to right feature Spearman rho correlation coefficients of 0.317 and 0.360 respectively, demonstrating statistically significant, but weak associations between the measures on each of the generic instruments (FFT or CLASS) and the math-specific one (MQI). None of the other six correlation coefficients are substantially different than the ones mentioned above, ranging from 0.25 to 0.42. The transition matrices reflect and better support these results as most of the teachers' quality rankings do not stay around the same place in the distribution. Most of the diagonal elements (including the two ends) are close to 10%, which is

the random assignment baseline. The transition matrix in Table 19 features a comparison pair with the lowest correlation ($\rho = 0.259$) among the set of eight combinations.

Table 19: Rankings of Year Two mathematics teachers' FFT Average scores and MQI Average scores: Percentage of Teachers

Decile Rank in MQI	Decile Rank in FFT (N = 770)									
	1	2	3	4	5	6	7	8	9	10
1	9.9%	20.8%	13.2%	10.2%	11.9%	6.7%	3.9%	10.0%	8.6%	5.7%
2	19.8%	13.0%	13.2%	11.4%	6.8%	7.9%	14.3%	6.3%	9.9%	7.1%
3	21.0%	14.3%	13.2%	9.1%	16.9%	10.1%	9.1%	6.3%	4.9%	8.6%
4	6.2%	6.5%	11.8%	9.1%	3.4%	4.5%	6.5%	10.0%	1.2%	5.7%
5	7.4%	13.0%	11.8%	6.8%	10.2%	12.4%	14.3%	8.8%	12.3%	12.9%
6	7.4%	5.2%	10.3%	10.2%	10.2%	11.2%	7.8%	5.0%	7.4%	4.3%
7	14.8%	11.7%	8.8%	21.6%	10.2%	14.6%	10.4%	15.0%	7.4%	12.9%
8	8.6%	6.5%	5.9%	9.1%	15.3%	10.1%	10.4%	13.8%	14.8%	5.7%
9	0.0%	3.9%	4.4%	11.4%	10.2%	11.2%	18.2%	12.5%	12.3%	15.7%
10	4.9%	5.2%	7.4%	1.1%	5.1%	11.2%	5.2%	12.5%	21.0%	21.4%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Note: FFT is the reference group.

In the above situation, only 36% of the 770 mathematics teachers stay in the same or adjacent deciles, but 64% of them move up or down at least two percentile groups in the distribution. The other comparisons regarding FFT or CLASS and MQI have slightly higher percentages of teachers remaining in the same and/or adjacent place, but the instability between measures are still similar to the example above. In contrast to the comparisons between the two generic instruments, these patterns demonstrate the larger degree of variability in mathematics teachers' observational measures across generic and subject-specific instruments, even though the instruments are measuring the quality of individual teachers using the same set of lessons.

6.4. Chapter Summary

To conclude, when rating and ranking mathematics teachers' lessons, generic instruments agree with each other more than they agree with the subject-specific instrument across the

examined contexts. Though the volatility of teachers' ranks is still non-trivial in the middle of the quality ranking distribution, there is much higher stability in the bottom and top percentile groups when the generic instruments are used for measurement. Such stability at the two ends of the distribution holds some promises for the current practices in many districts, as the evaluation outcomes and retention/compensation decisions made with regards to the top and low performing teachers are consistent when different generic instruments are employed. Stakeholders should attend more to the type of an instrument when choosing a tool for implementation in the district, as there is a large variability in teachers' quality measures across generic instruments and subject-specific ones. Depending on the type of instrument selected, the rankings of teachers will look substantially different, hence the inferences with regard to teacher quality of the population will change accordingly.

CHAPTER 7

DIFFERENCES IN MATHEMATICS TEACHERS' OBSERVATIONAL RATINGS ACROSS SUBJECT AREAS

7.1. Introduction

This chapter addresses the third research question: For the mathematics teachers who taught diverse mathematical topics, to what extent are their observational scores different across subject areas as measured by various protocols in the MET data? Prior literature has discussed the nuances in the knowledge base for teaching different mathematical content (Groth, 2007; McCrory et al., 2012), but whether the nuances in teaching practices translate into teachers' observational ratings was not examined. In teacher evaluation, if there was evidence that supported the influence of subject area on mathematics teachers' observational ratings, the implicitly held assumption that there is a unifying teacher quality or qualities across areas of mathematics for the same teacher should be challenged. If mathematics teachers do not get stable observational ratings when they are teaching lessons with different content, neglecting the subject matter of the observations might return a biased estimate of the average teacher quality across lessons, and accordingly might return inappropriate and unfair evaluation results. In other words, the validity of the inference a stakeholder makes with regard to teachers is at stake if one does not differentiate the content of the mathematics lessons from which the scores are attained.

To examine the differences in teachers' observational ratings across subject areas within mathematics, I group mathematics teachers into several samples based on the content of the four observations collected from them[27], and compare the same teachers' ratings in one subject area

[27] As discussed in Chapter 3, considering the sample size, the pairs of subject areas examined are: 1) Algebra & Algebraic Thinking (AA) vs. Numbers & Operations (NO); 2) Numbers &

to another within each instrument. Similarly to how I answer the first research question in Chapter 5, the comparisons are made and evaluated around three parameters that are relevant in teacher evaluation systems: 1) whether the choice of the instrument matters, 2) whether the score aggregation methods (PCA vs. simple average) matter, and 3) whether the adoption of unlike frameworks to categorize teachers (absolute performance vs. relative performance) matters.

7.2. The Influence of Subject Areas on Mathematics Teachers' Observational Raw Scores

Analyses using paired-sample t-tests of the MET data show that the subject area of the lessons does not have substantial influence on mathematics teachers' observational ratings. Of the 60 comparison pairs[28] involving different observational protocols and score generation methods, mathematics teachers' scores based on observations in one area of mathematics are not significantly different from those in another area for 54 combinations. For the few comparison pairs that have significant results, despite the fact that their difference cannot be explained by the grade level factor and the district factor[29], the discrepancy only shows up in one year and does not persist in both years to support the prevalence of such a difference. Table 20 summarizes the P-values and effect sizes of those statistical significant comparison pairs and their non-significant counterparts in the other year, if any.

Operations (NO) vs. Geometry (G), and 3) Algebra & Algebraic Thinking (AA) vs. Statistics & Probability (SP).

[28] For comparison pairs examined to answer the third research question, please see Appendix D.

[29] For the ANOVA models examining the main effect and the interaction effect involving subject as well as grade level/district with regard to the mathematics teachers, please refer to Appendix J.

Table 20: P-values and effect sizes for significant comparisons and the insignificant counterparts in the other year

Instrument	Subject Areas	Comparison Level	Year One	Year Two
FFT	NO vs. G (N = 135, 84)	Management	0.453	0.041* (0.227)
CLASS	AA vs. NO (N = 231, 175)	Support	0.033* (0.141)	0.921
		Average	0.046*(0.132)	0.861
MQI	NO vs. G (N = 125, 81)	Accuracy	0.005** (0.249)	0.282
		Average	0.030* (0.194)	0.812
	AA vs. SP (N = 55)	Accuracy	0.032* (0.289)	

Note: * means the difference is significant at the 0.05 level. ** means that the difference is significant at the 0.01 level. If not significant at least at the 0.05 level.
The bolded subject areas with underscores are the ones that receive higher ratings in the significant cases.

For the above significance cases, if adjusting the cut-off p-value based on the Bonferroni approach within instrument by year to control for the probability that these results merely show up by chance, none of the comparisons are significant except for one pair. The pair of Numbers & Operations and Geometry is significant in the MQI ***Accuracy*** component with the Year One sample ($p = 0.005$). This particular significance suggests that teachers made fewer errors and were more precise when they were teaching Numbers & Operations lessons than teaching Geometry lessons across grade levels, as reflected by their MQI scores. But in other measures of teacher quality from various instruments, there is no sufficient evidence to say Numbers & Operations lessons receive better quality ratings than Geometry in general when taught by the same teacher.

In conclusion, the content of a mathematics lesson is not the deciding factor of mathematic teachers' observational ratings and evaluation results, as 90% of the time the same teachers'

scores are undifferentiated across content. This suggests a stable estimate of the unifying mathematics teacher quality or qualities across subject areas from those observational protocols examined.

7.3. The Influence of Subject Areas on Mathematics Teachers' Observational Rank Scores

Despite the lack of variations in raw scores between pairs of subject areas at the teacher level, the rank scores of these mathematics teachers' observations suggest that teachers do not have similar quality rankings when they are observed and evaluated in dissimilar mathematical content, and hence return inconsistent results in evaluation under the rank-order tournament framework. The significant different cases in raw scores are the ones that do not demonstrate stability across the whole score distribution, even in the bottom percentile groups where teachers' retention is at stake. Among the other non-significant comparison pairs, teachers' quality rankings across areas of mathematics are more consistent when teachers' lessons are assessed in the generic instruments than in the math-specific one, especially in the bottom percentiles.

The transition matrices below are two typical examples to demonstrate teachers' quality rankings between pairs of subject areas in the different types of observational protocols. Table 21 shows a transition matrix for a comparison pair on MQI that has a significant difference. In contrast, Table 22 shows a transition matrix for one of the 54 non-significant comparisons whose scores are attained from a generic instrument—CLASS.

Table 21: Rankings of Year One mathematics teachers' MQI Accuracy scores of NO vs. G: Percentage of Teachers

Decile Rank in G	Decile Rank in NO (N=221)									
	1	2	3	4	5	6	7	8	9	10
1	8.3%	7.7%	16.7%	0.0%	15.4%	17.6%	0.0%	16.7%	0.0%	8.3%
2	33.3%	15.4%	8.3%	16.7%	7.7%	5.9%	11.1%	0.0%	0.0%	8.3%
3	0.0%	15.4%	25.0%	0.0%	7.7%	11.8%	22.2%	8.3%	7.7%	8.3%
4	25.0%	7.7%	8.3%	8.3%	7.7%	17.6%	11.1%	0.0%	7.7%	16.7%
5	0.0%	15.4%	0.0%	0.0%	15.4%	5.9%	0.0%	16.7%	7.7%	16.7%
6	16.7%	7.7%	8.3%	25.0%	15.4%	11.8%	11.1%	0.0%	7.7%	0.0%
7	0.0%	15.4%	25.0%	25.0%	7.7%	5.9%	33.3%	0.0%	15.4%	0.0%
8	0.0%	15.4%	0.0%	8.3%	0.0%	11.8%	0.0%	8.3%	15.4%	16.7%
9	16.7%	0.0%	0.0%	8.3%	15.4%	5.9%	11.1%	16.7%	15.4%	16.7%
10	0.0%	0.0%	8.3%	8.3%	7.7%	5.9%	0.0%	33.3%	23.1%	8.3%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Note: Numbers & Operations is the reference group.

Table 22: Rankings of Year Two mathematics teachers' CLASS Average scores of AA vs. NO: Percentage of Teachers

Decile Rank in NO	Decile Rank in AA (N=175)									
	1	2	3	4	5	6	7	8	9	10
1	52.9%	16.7%	27.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
2	5.9%	33.3%	11.1%	0.0%	15.0%	10.5%	12.5%	6.3%	4.8%	0.0%
3	17.6%	0.0%	11.1%	14.3%	5.0%	26.3%	12.5%	6.3%	4.8%	0.0%
4	0.0%	11.1%	16.7%	7.1%	20.0%	5.3%	18.8%	12.5%	4.8%	6.3%
5	11.8%	11.1%	11.1%	14.3%	15.0%	21.1%	6.3%	0.0%	4.8%	6.3%
6	5.9%	0.0%	0.0%	7.1%	15.0%	5.3%	12.5%	18.8%	9.5%	12.5%
7	0.0%	5.6%	11.1%	21.4%	5.0%	5.3%	12.5%	12.5%	9.5%	25.0%
8	0.0%	11.1%	5.6%	7.1%	5.0%	10.5%	0.0%	6.3%	23.8%	18.8%
9	5.9%	5.6%	0.0%	21.4%	10.0%	10.5%	6.3%	18.8%	23.8%	18.8%
10	0.0%	5.6%	5.6%	7.1%	10.0%	5.3%	18.8%	18.8%	14.3%	12.5%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Note: Algebra & Algebraic Thinking is the reference group.

In the first transition matrix above (Table 21), 34.4% of the teachers stay in the same percentile groups or the adjacent ones as evaluated in different areas of mathematics by a

subject-specific instrument—MQI. Moreover, of the 221 teachers in the sample, 65.6% of them move up or down at least two or more percentile groups when evaluated in another area of mathematics.

In comparison, the second transition matrix (Table 22) features higher percentages along the diagonal than the first one, especially at the two ends of the distribution. Overall, 42.9% of the teachers stay in the same percentile groups or the adjacent ones when evaluated in different subject areas by CLASS. Additionally, 57.1% of the 175 teachers move up or down at least two or more percentile groups when evaluated in another area of mathematics. Nonetheless, the high percentages for teachers who changed at least two percentile groups in both cases are alarming, as it signals that the majority of the mathematics teachers would have teacher evaluation results one or more quality category higher or lower when they are rated and ranked separately for different subject areas they teach.

The phenomenon of low percentages along the diagonal cells and their adjacent ones prevail for all comparison pairs, regardless of what aspect of teacher quality one attends to and which score aggregation algorithm is used. To conclude, mathematics teachers' quality rankings based on different mathematical areas volatilize greatly within the population, especially for the mathematics teachers who have middle rankings and/or are evaluated by a subject-specific instrument. Such instability in the same teachers' rankings hence returns inconsistent teacher evaluation results when one is observed in certain areas of mathematics.

7.4. Chapter Summary

In this chapter, I investigate several perspectives that are relevant for teacher evaluation practices to compare mathematics teachers' ratings across various areas within the discipline. It turns out that depending on which performance framework is used, mathematics teachers'

observational ratings may or may not be decidedly different when they are observed in dissimilar content. One possible explanation is that mathematics teachers' observational scores generally lack variations and are similar in absolute values. In that case, under the fixed performance model where certain thresholds were set up to categorize teachers, the majority of them would be clustered in one or two quality categories. This is actually the controversial state-of-the-field in many teacher evaluation systems, as almost all teachers are in the middle quality category. Under the rank-order tournament model, however, even the slightest difference in raw scores will impact ranking. Accordingly, the results in this chapter show that mathematics teachers' rank scores from pairs of subject areas vary greatly within the population, and they are even more volatile when measured by the math-specific instrument than by the generic ones. In the end, one should still caution against neglecting the role played by content of mathematics teachers' observations in given contexts.

CHAPTER 8

CONCLUSION AND DISCUSSION

8.1. Summary of Findings and Discussion

The primary aim of this study is to investigate how the content of the lesson, as well as other contextual factors (the choice of an instrument, ways to generate composite scores, frameworks to categorize teachers) can affect observational ratings of teachers, and hence may be consequential for teachers' evaluation results. Following Kane's (2006, 2012, 2013) argument-based approach framework and his notion of construct validation, this study adds to the research program of the teacher quality construct by examining the assumptions, interpretations, and consequences of using observational scores attained from a variety of instruments, specifically for the purpose of teacher evaluation. The characteristics of the MET data allow for an examination of the those potentially influential factors as the same group of teachers was rated by various instruments at multiple times, which is unrealistic to replicate in actual teacher evaluation. Accordingly, the results of this study can be of theoretical as well as practical use for both researchers and policymakers who are interested in measuring teacher quality for different purposes, even going beyond teacher evaluation.

8.1.1. Results of Generalist Teachers

One of the main findings of this study is that subject does matter for generalist teachers' evaluation results, as discussed in Chapter 5. Prior research has established the differences in the cultures (Grossman & Stodolsky, 1995; Stodolsky & Grossman, 1995) and knowledge for teaching ELA and mathematics (Ball, Thames, & Phelps, 2008; Grossman, 1990; Hill, et al., 2008; Shulman, 1987). The results of this study also show that such differences translate into teachers' differentiated scores across a variety of contexts, including various instruments, score

aggregation methods, incentive structure models in gauging performance, and grade levels. In particular, for three-fourths or more of the comparisons, the same teachers' ELA and mathematics ratings differ in both the absolute value and the rank scores. This holds true whether the evaluator uses a generic instrument or a subject-specific one, and whether the evaluator uses the component scores or the simple average scores to represent the teacher's quality/qualities. In particular, in the case of CLASS, teachers' ELA scores are substantially higher than their mathematics ones in the managerial aspect of teacher quality, and this large discrepancy prevails across elementary grade levels. Given this discrepancy, consider an evaluator walking into a teacher's mathematics lessons rather than the ELA ones; he or she would on average rate the teacher two standard deviations lower in raw scores on dimensions related to classroom ***Organization*** in CLASS.

Furthermore, under the rank-order tournament framework, when generalist teachers were assessed and ranked in both subjects by generic instruments, they are more likely to be placed in the same and/or adjacent location in the quality ranking distribution within the same group of teachers. There could be several reasons for this phenomenon. One hypothesis is that using generic instruments to assess lessons of both subjects tends to capture more of the commonality of the teacher quality construct across content, in comparison to using the subject-specific instruments for ELA and mathematics respectively. If the purpose of teacher evaluation is to provide a more stable estimate of the unifying teacher quality or qualities shared by the same teachers' observations across different subjects, the generic instruments perform better than the subject-specific ones.

8.1.2. Results of Mathematics Teachers

In Chapter 6, the comparisons of mathematics teachers' rank scores across combinations of observational instruments show that the variability of their observational ratings is larger when comparing generic to subject-specific instruments than when comparing across generic instruments. In other words, the ratings on generic instruments (FFT and CLASS in question) are more consistent with each other than with the ratings on a math-specific instrument (MQI). One explanation for this result is that there is a large overlap in the construct of teacher quality/qualities that both generic instruments target to measure. None of the existing literature on classroom research and reports on teacher evaluation have explicitly evaluated the utility of generic and subject-specific instruments for the purposes they are used for, and/or discussed the affordances and constraints of using a particular type of observational protocol. The results of Chapter 6 suggest that the unconditional use of certain types of instruments may result in different inferences and consequences. Practically, within teacher evaluation systems, stakeholders should bear in mind that depending on the instrument type used for observations, mathematics teachers' evaluation results are subject to a large shuffle. It is hence important for stakeholders to clearly define the targeting goals and criteria for the instrument selected.

Chapter 7 presents the results of whether the particular content being taught is associated with differences in the same mathematics teachers' observational ratings. In contrast to the findings regarding the importance of subject (ELA vs. math) in generalist teachers' observational raw scores, mathematics teachers' ratings do not depend on the mathematical content of the lessons. That is, some of the noted nuances in knowledge base and practices across areas of mathematics (e.g., Herbst & Kosko, 2014; Groth, 2007; McCrory, et al., 2012) do not translate

into mathematics teachers' observational raw scores across instruments and score aggregation methods. Overall, 90% of the comparisons between pairs of subject areas do not show statistical significance, indicating the stability of measures across subject areas for the same teacher. For the 6 out of 60 total cases that are significant, the patterns only emerge for one year of analysis but are not present for the other, making the likelihood of differences less robust due to the lack of repetition of significance with another sample. For example, mathematics teachers from the Year One sample demonstrate differentiated scores that are significant between Numbers & Operations and Algebra & Algebraic Thinking as measured by CLASS ***Support*** component scores as well as its simple average scores. But mathematics teachers from the Year Two sample (with a large overlap with the Year One sample) drawn from the same districts do not demonstrate such patterns. One reasonable explanation is that this occurrence is measurement error: as raters in the study got more experienced scoring, they might be more reliable raters, or the significance only happens by chance with this sample. Another alternative explanation is that there may be some forms of professional development intervention related to Numbers & Operations. The intervention may be targeting instructional and emotional support moves in the classroom, which may intentionally or unintentionally close up the gap between the teaching practices of these two areas of mathematics.

Another result presented in Chapter 7 tells a different story about the role played by the content of the observations in mathematics teachers' ratings. Under the rank-order tournament framework where teachers are ranked against each other within the same population, the subject area of the mathematics lessons in which teachers were observed are essential to decide where teachers are placed in the score distribution. Such influence of the content is prevailing across all comparison pairs, including the different instruments and the score aggregation algorithms.

Additionally, similar patterns emerge on the performance of the generic and subject-specific instruments discussed above. Generic instruments provide more stable estimates of the average teacher quality or qualities in teaching mathematics across its subject areas, especially for the highest and lowest performing teachers in all the comparisons examined. If the purpose of the teacher evaluation is to capture the commonality of mathematics teachers' quality/qualities across different subject areas and provide more consistent estimates, once again the generic instruments perform better than the math-specific one.

The lack of dispersion in mathematics teachers' raw scores might be the reason why the tests are not significant across different pairs of subject areas within mathematics. There are two hypotheses that may contribute to this observed phenomenon despite the conceptualized disparities in knowledge and practices across areas of mathematics. The first hypothesis is that, there may not be much subject-area-specific teaching practices for mathematics in enactment; teachers may tend to teach all kinds of mathematical topics with the same repertoire, or they were just under-prepared to do so in teacher preparation. The second hypothesis is that none of the instruments conceptualize instruction at the level of subject areas, which is more of a theoretical challenge than anything else. It may be easier to theorize instruction that is more extreme, as many teacher educators and administrators can recognize a brilliant or poor performance instantly when they see one. But it is difficult for them to come to consensus and to evaluate what is in between the two spectrums of "good" and "bad" instruction in mathematics, let alone in specific areas of mathematics. The prevailing instability between teachers' measures in the middle of the quality ranking distribution in all comparison pairs confirms the persistence of such theoretical challenges facing the research community.

8.2. Implications for Educational Policy and District Stakeholders

My analysis suggests that it is unfair and inappropriate to compare a group of generalist teachers' quality measures solely from their ELA instruction to another group of generalist teachers' quality measures solely from their mathematics instruction. Given the influence of subject on generalist teachers' observational ratings and evaluation results, if a teacher evaluation system wants to accurately and reliably represent the average quality/qualities of a generalist teacher, it should consider the content of the lessons being observed and collect the information for use in a systematic way. Purposeful sampling to attain a balanced set of observations from both subjects for each teacher is an important practice for the stakeholders to consider. Alternatively, stakeholders in districts may also consider separating teachers' evaluation results by subject given that the numbers of teacher who demonstrated large discrepancy in their instructional quality measures across subjects are non-trivial. By separating teachers' evaluation results by ELA and mathematics, districts can make strategic decisions about resource targeting in order to help those teachers with huge gaps between their ELA and mathematics instruction.

For mathematics teachers, depending on the framework under which the scores are used, the content of the lessons may or may not have apparent consequences. This study does not find sufficient evidence to support separating the teacher evaluation results by areas of mathematics. But still, since there is some effect of the content from a small numbers of comparison and under particular contexts, purposeful sampling can help minimize the biases toward certain content when capturing the average quality/qualities of a mathematics teacher. Most, if not all, observational protocols do not have specifications about the content that is at the grain size of the subject area. But recording the content of a mathematics lesson and/or intentionally selecting a

mathematics lesson of certain content domain is not a costly practice for stakeholders to consider implementing at scale, and this practice is beneficial for districts to provide supports to teachers around particular content domain within mathematics.

The second implication for education policymakers is regarding the score aggregation methods. There are four ways to generate observational scores from dimensions: 1) taking simple averages across all dimensions; 2) taking averages within domains that are initially built into the instrument, 3) attaining multiple component scores from factor analysis, and 4) taking averages of weighted dimensions. The first and the third methods are examined in this study as the conditions for one of the three relevant parameters in teacher evaluation. The second method is emerging from the results and some of the practical considerations. The last method is not examined in this study but has been explored by other researchers in the field of educational policy research.

Based on the results of this study and considering the affordances and challenges of these practices, I would recommend the districts to adopt the method of taking averages within clusters of similar dimensions (i.e., domains) to represent individual teachers' qualities. The recommendation is made for the following reasons. First, this method helps analyze teachers' qualities at a finer grain size while efficiently attaining individual teachers' observational scores at a large scale. Also, based on the PCA results in Chapter 4, since the factor patterns of the MET data roughly confirm the domain structure that developers of FFT, CLASS, and PLATO[30] initially built into the instruments, this method is as good as the complicated factor analysis in terms of the information retained. Last but not least, this method provides more information than

[30] For MQI-Lite, the dimensions used in the MET study is equivalent to the domain in other instruments, as there are many sub-dimensions pertaining to the dimensions in the MQI-Full version, but not used in the MET study to rate the lessons.

the simple averages across all dimensions within an instrument by reflecting teacher qualities at the sub-construct levels.

The other methods each have some affordances and constraints for the districts to adopt, depending on their needs and preferences. For the PCA algorithm, despite the cost to implement this practice at a large scale, I argue that there is still valuable information in teachers' observational ratings on particular aspects of teacher quality generated from factor analysis, especially in comparison to the simple averages. In this study, even though calculating component scores takes more efforts than averaging across all dimensions, the simple average scores sometimes can cause information to confound among the different aspects of the teacher quality construct. For example, with the Year Two generalist teacher sample evaluated by CLASS, the ***Support*** aspect of the teacher quality as well as the simple average both show that mathematics lessons have slightly but not significantly higher ratings than ELA ones for the same teachers, but the ***Organization*** component scores indicate teachers' ELA lessons receive ratings that are two standard deviations higher than their mathematics ones. If the districts only knew about teachers' simple average ratings on CLASS, they would learn the same thing from the simple averages as from the component scores on the instructional support aspect of the classroom processes. Yet at the same time, the average scores also offset the large differences in the managerial aspect that tell an opposite story about teachers' performance in ELA versus in mathematics—this discrepancy may be of great interest for feedback and resource allocation.

For stakeholders who prefer a single composite score at the teacher level for policy, an average with various weights depending on the instrument is an alternative method to generating component scores from factor analysis or attaining domain-level averages. The advantage of using averages with weights is that districts can then use the single composite at the teacher level

to represent teacher quality and make more straight-forward decisions regarding tenure, compensation, and retention rather than dealing with multiple scores for each teacher. At the same time, this single composite accounts for the content-dependent nature of teachers' ratings, and include more information on the different aspects of the teacher qualities that is lost with the simple averages. The question then becomes, how do we take the differentiated scores that a teacher attains from being observed in teaching a variety of content in order to generate a meaningful and valid single composite? There is no readily available answer to this question. Past research on how to combine different measures of teacher quality (including the observational measures) into a single composite measure can shed some light on this matter (Mihaly, McCaffrey, Staiger, & Lockwood, 2013). Following prior literature and evidence, to generate a single composite score one needs to empirically explore competing approaches, specify the underlying concept to be measured (e.g., high leverage practices), and target criteria (e.g., state policies). Different methods of combining teachers' ELA scores and mathematics scores from diverse instruments should be explored by comparing and contrasting various weighted composites. The weighted composites can be generated with equal weights (which is the simple average by definition), or they can be generated with statistical weights that make them the optimal predictor of a set of target criteria decided by the district. The set of target criteria may include students' achievement data in both ELA and mathematics respectively, as well as survey data that also differentiated students' learning experiences across content in order to account for the influence of subjects. Although this study does not examine the utility of such method, it still adds to the argument from this line of research that the methods to combine scores from lessons of different content depend on the instrument used, but finds no evidence on the impact introduced by grade levels.

8.3. Limitations of the Study and Suggestions for Future Research

First, there are some differences in terms of the modes to collect observational data between the MET study and the actual teacher evaluation processes. The MET project features video scoring to capture the teacher quality. A large group of well-trained raters were assigned to score videos in certain subject, and each of them only focus on a small set of dimensions within each instrument in order to avoid cognitive overload. This is a huge effort to ensure reliability, but such practices are not adopted in the actual teacher evaluation systems, nor are they possible because of the cost in personnel and time. Teacher evaluation systems tend to have a small number of administrators and expert teachers coming to teachers' classroom and conduct live observations in order to evaluate teachers' quality, sometimes the evaluator may not even be a teacher in the subject that is being observed. One noted difference is on attaining teachers' observational scores from doing live observation versus from watching a video. As Casabianca et al. (2013) have shown in their study, there is evidence that scores from live observations tend to be higher than from videos with the particular observational protocol examined. We do not know, however, whether the discrepancy between ratings from live observations and videos are consistent, nor do we know whether such difference persist when other observational protocols are used. Accordingly, the results of this study on the influence of content and other contextual factors may not be completely generalizable to the implementation of teacher evaluation as the observation mode changes. Future studies can examine the observational ratings attained from live versus from video systematically across instruments to support the generalizability of the results in this paper.

Second, this study only examines the observational scores without examining their connections with the student learning outcomes. Future studies may use the PCA results and the

component scores to see which component significantly predicts student learning outcomes as measured in the Value-added Model (VAM). Because the components are unrelated to each other within an instrument by construction, there is no issue on co-linearity between the independent variables entered in the model. If a certain component is not significant in predicting student learning outcomes, teacher evaluation can consider lowering the weight of this aspect of teacher quality when evaluating teachers given that teacher evaluation is premised on improving student achievement and college readiness.

Moreover, as much as I tried to compare the MQI-Lite in this study with other generic instruments to understand how they differ in practice on top of the conceptual and methodological disparities, with the restricted set of dimensions in MQI-Lite, there is no convincing evidence on how a math-specific instrument and the generic instruments operationalize differently to measure instruction, and more specifically, mathematical subject area instruction. Even though PLATO does not have limited dimensions as MQI-Lite for ELA instruction, for the focus of this study, I did not compare the generic instruments to the ELA-specific one to support or caution against the prevailing advocate of using subject-specific protocol. Future research can use the MQI Full/PLATO or other subject-specific instruments to compare with the generic ones, and examine the affordances and challenges of using certain tools from the perspective of research on teaching and learning, the perspective of professional development, as well as the perspective of teacher evaluation.

Additionally, this study only compares rank scores when it comes to cross-instrumental comparisons of the same teacher. More can be done with the use of observational score equating before comparing the differences across instruments on the same pairs of conceptually matched

teacher quality components or the simple averages for the same teacher. Such methodology includes item response theory, equipercentile equating, etc.

Last but not least, further research is needed to determine whether the differences associated with content are due to the capacity of an instrument, or due to the real variations in teachers' quality and instructional practices across content. Many of the significant differences detected in this study are in the managerial-related components measured by the generic instruments (FFT ***Management*** and CLASS ***Organization***). As implied by teachers' higher ratings in ELA, they seem to be able to manage their ELA lessons in a more organized way than their mathematics ones, even though they are equipped with all necessary teaching repertoire. Accordingly, what contributes to the changes in teacher's instructional moves in terms of classroom management across subjects is a topic meriting further investigation. Qualitative studies with videos can help educators understand what may contribute to the differences in terms of classroom organization/management, especially between ELA and mathematics lessons. Such qualitative studies have the potential to bridge pedagogical content knowledge and general pedagogical knowledge conceptually and empirically.

Another component with occasional differences detected is the stand-alone component—MQI ***Accuracy***. Qualitative studies focusing on teachers' errors and imprecisions provide insights into teacher knowledge and practices for teacher educators. For example, there is a significant difference in terms of accuracy between the same teacher's teaching in Algebra & Algebraic Thinking lessons and Statistics & Probability ones. It is interesting to analyze the errors and imprecisions that in-service teachers have, and to understand the reasons why teachers tend to have more errors and imprecisions in certain subject areas than others. Such efforts can

be capitalized in our teacher preparation and/or other ongoing professional development programs in order to address the gaps in teachers' mathematical knowledge for teaching.

One possible direction to go for the future research is to examine the instructional practices of those teachers who demonstrate the largest discrepancy in their teacher quality measures across content. What makes an elementary teacher a top performing teacher in ELA but at the same time the poorest mathematics teacher? What makes a teacher a top performing teacher in algebra but at the same time the poorest teacher in statistics? What does it mean to be a good elementary teacher or a good mathematics teacher? The results of this study provide the potential samples to start the investigations of these questions empirically, and help educators better theorize the good and bad instruction that is situated in the content of our school curriculum.

APPENDICES

APPENDIX A:

STATISTICS AND PROBABILITY LESSONS AND THEIR CONTENT IN THE MET VIDEO DATA

Table A 1: Grade 6 Statistics & Probability lessons (N = 36)

ID	Topic	Year	Original Topic in Video Information File
aar7	Frequency Table	2	Random Topic
aga7	Frequency Table	1	Random Topic
ari4	Median & graph	2	Random Topic
baw9	Probability	1	Random Topic
bdv2	Mean & calculations	2	Random Topic
bgg5	Reading graphs	1	Random Topic
cdz5	Probability	1	Random Topic
cie4	Mean, mode, median and range	2	Random Topic
ckd2	Mean, mode, median and range	1	Random Topic
cmt2	Sample space, circle graphs, making prediction, and probability	1	Random Topic
cnt7	Mean, mode, median and range	1	Random Topic
crk8	Probability (permutation and combination)	1	Random Topic
crs4	Probability	2	Random Topic
cuc5	Box plot, and measures of central tendency	2	Using the commutative, associative, identity and distributive properties
czy2	Line plot for data	2	Creating, analyzing tables, graphs and equations to describe linear functions and other relationships
dbd2	Line graph , bar graph and frequency table	1	Random Topic
dbe4	Sample space	1	Random Topic
dex3	Outcome of compound events	2	Creating, analyzing tables, graphs and equations to describe linear functions and other relationships
dfd4	Probability	1	Random Topic
dfk2	Mean, mode, and median	2	Random Topic
dhq4	Line graph, bar graph and frequency table	2	Creating, analyzing tables, graphs and equations to describe linear functions and other relationships
dwg5	Line graph frequency table, stem and leaf plots, histogram, bar graph, etc.	2	Creating, analyzing tables, graphs and equations to describe linear functions and other relationships
edx2	Theoretical and experimental probability	2	Random Topic
Egg8	Probability	2	Random Topic

Table A 1 (cont'd)

ID	Topic	Year	Original Topic in Video Information File
fmy3	Intro to statistics (what is data, how are data collected, etc.)	1	Random Topic
fpm4	Theoretical probability	1	Random Topic
ftz9	Probability	2	Random Topic
fwv4	Probability	2	Random Topic
gak4	Probability of compound event	1	Random Topic
gbd3	Probability	1	Random Topic
gqk8	Line graph, frequency table, mean, median, mode, and range	2	Random Topic
gui8	Experimental probability	1	Random Topic
gvc7	Mean	2	Random Topic
hkq7	Mean, mode, median and range	1	Random Topic
hps2	Combinations (outcomes)	2	Random Topic
hrb5	Probability	1	Random Topic

Table A 2: Grade 7 Statistics & Probability lessons (N = 32)

ID	Topic	Year	Original Topic in Video Information File
aqt3	Mean, median, mode, and range	1	Random Topic
awz5	Probability	2	Random Topic
bcn9	Mean, median, and mode	1	Random Topic
bie5	Stem and leaf plot	2	Random Topic
bkk3	Data collection, survey (bias)	2	Random Topic
cbh3	Mean, median, mode, and range; quartile	1	Random Topic
cuh2	Bias in survey	2	Random Topic
cxf2	Line graph, bar graph, circle graph, and picture graph	2	Random Topic
dar6	Stem and leaf plot, bar graphs, circle graphs (analyzing data)	2	Random Topic
ddt9	Box plot	2	Random Topic
des8	Data, outlier, histogram	1	Random Topic
dhb2	Biased and unbiased sample	2	Random Topic
dqr8	Biased and unbiased sample	2	Random Topic
ebv2	Probability	2	Random Topic
emq5	Probability	2	Random Topic
ene9	Probability	2	Random Topic
esm9	Histogram and frequency table	2	Random Topic
eux5	Mean, median, and quartile	2	Random Topic
fkn7	Analyzing and constructing circle graphs	2	Random Topic

Table A 2 (cont'd)

ID	Topic	Year	Original Topic in Video Information File
fkt7	Select appropriate measure of central tendency (mean, median and mode)	1	Writing, interpreting, and/or using mathematical expressions and equations
fpr7	Line graph, bar graph, circle graph, steam and leaf plot and box plot	2	Linear Equations
fvt3	Box plot	2	Random Topic
fwu6	Experimental probability	2	Random Topic
gbv6	Probability, event, outcome	1	Ratio, rate, and proportional reasoning
bcn5	Biased and unbiased sample	2	Cannot find in rated video information file
bdc2	Probability, event, outcome	2	Cannot find in rated video information file
gfy5	Measure of variation	2	Random Topic
gui9	Probability	2	Random Topic
hbp9	Mean, median, mode, range, outlier, and box plot	1	Random Topic
hnb7	Probability	1	Random Topic
hqp4	Mean, median, mode, histogram	1	Random Topic

Table A 3: Grade 8 Statistics & Probability lessons (N = 16)

ID	Topic	Year	Original Topic in Video Information File
auw3	Sample vs. census	1	Random Topic
bpc3	Box plot, quartiles, and independent events	2	Random Topic
cfg2	Box plot, theoretical and experimental probability	2	Random Topic
ctg6	Various representations: circle graph, Venn diagram, stem and leaf plot, and box plot	1	Random Topic
cwu6	Measure of central tendency	2	Random Topic
dmf9	Box plot	2	Random Topic
dpt8	Mean, median, mode and range	1	Random Topic
dwc7	Measure of variation, box plot	1	Random Topic
ekc5	Probability of composite experiment	2	Random Topic
epm4	Theoretical probability	1	Random Topic
euz4	Independent and dependent event, probability of event with replacement and without replacement	2	Rate, ratio, and proportional reasoning
fny7	Experimental probability	1	Random Topic
hat3	Box plot	2	Random Topic
hem7	Reading graphs	2	Rate, ratio, and proportional reasoning

Table A 3 (cont'd)

ID	Topic	Year	Original Topic in Video Information File
hiq8	Probability	2	Random Topic
hkp7	Median, box plot, stem and leaf plot	2	Random Topic

APPENDIX B:

PRINCIPAL COMPONENT ANALYSIS PROCESSES

B.1. Introduction to Principal Component Analysis

The four instruments, similarly to most other observational instruments out there currently, feature multiple dimensions that are correlated with each other to a medium extent round 0.4. The purpose of PCA is to answer three questions: 1) how many aspects (components) of teacher quality do these observational instruments really measure; 2) what are those aspects (components) and their meanings; and 3) to what extend does each dimension (variable) measure each of these aspects (components).

Generally speaking, there are four steps for factor analysis, including PCA. First, use the variables to compute the correlation or covariance matrix, and compute statistics on the factorability of the matrix. Second, extract an initial solution and determine the numbers of factors/components to keep in the final solution. Third, examine the meaning of the factors/components, and rotate them if necessary to clarify the data pattern in order to better interpret the nature of factors/components. Lastly, use factor/component loadings to calculate factor/component scores for each participant for any subsequent applications of the results from data reduction.

B.2. Suitability to Conduct PCA

In my analyses, I computed the correlation matrix in each case because variables within each instrument are on the same scale, and do not have very different means and standard deviations. There are two common tests used by statistical software packages to demonstrate the factorability of the correlation matrix: Barlett's test of sphericity and Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (Cerny & Kaiser, 1977; Kaiser, 1970). The null

hypothesis of the Barlett's test is that the correlation matrix is an identity matrix, which means that the variables are not collinear, and the non-zero correlations in the sample matrix are due to sample error. To justify the use of PCA on the data, the null hypothesis needs to be rejected at the significance level of 0.05. The PCA I did with subsamples and full samples on all four instruments have passed this statistical test, suggesting some extent of collinearity among variables. The KMO test is to see to what extent can the amount of variance in the data be accounted by the extracted components. Value of 1 means all the variance can be accounted for by the extracted components. Kaiser et al. as well as other statisticians have suggested that KMO value higher than 0.6 as the threshold. In the different iterations of PCA for each instrument, the analyses with FFT and CLASS have high KMO statistics (>0.9 in general). But for subject-specific instruments—MQI and PLATO—the KMO statistics are at the 0.6~0.7 range. The factorability is still acceptable and a fair amount of variance is explained by the extracted components, even though not a substantial amount.

B.3. Criteria Used in PCA

There are two criteria that people use in order to decide the numbers of component to keep as sufficient representation of the dimensionality of the data. One criterion is called the Kaiser criteria, which is to keep components with corresponding eigenvalue larger than one. The larger the eigenvalue, the larger percentage of variance out of the total variance is accounted for by the corresponding component. The other criterion is to see whether the accumulative variance explained has exceeded certain threshold, usually 60% of the total variance. Many times the two criteria can be met at the same time, but sometimes only one of them are met and the other is close to be met. In some of the PCA I did in Chapter 4, I encountered situations where the eigenvalue is very close to, but did not exceed one. But without this corresponding component,

the total variance explained are below 60%. In such cases, I forced in the component with close-to-1 eigenvalue so that the total variance explained in each instrument's PCA is higher than 60%. In other words, I prioritized the second criterion when it comes to decide the numbers of component to keep for calculation of component scores later.

After the components are extracted, one will observe the coefficient of the variables—elements in the eigenvectors—within each component to understand what exactly the component measures. For the coefficient, the higher it is in absolute value within the component, the higher it correlates with the component. The naming of the component is based on the meaning and interpretations of these highly component-correlated variables. Many times, the components in the initial solution may be difficult to interpret because many variables' loadings are relatively high within each component, and some variables might have relatively high absolute loadings across different components. In order to better interpret the nature of the components extracted, rotation of component is often performed to help clarify the factor pattern. In this study, I used orthogonal rotation methods to keep components independent of each other after rotation. Moreover, I employed three orthogonal rotation methods with fixed gamma parameter—Varimax, Equimax, and Quartimax—to see which one improve the patten the best. Varimax maximizes the sum of the variances of the squared loadings. That is, in each component, the large loadings are increased and the small ones are decreased so that each component only has a few variables with large loadings. Equimax rotates the loadings so that a variable loads high on one component and loads low on other components. Quartimx maximizes the variance of the squared factor loadings in each variable. That is, for each variable, the large loadings are increased and the small ones are decreased so that each variable only loads on a few components. In this study, I prioritized the goal of having larger difference between higher loadings and lower

loadings within each component when I examined the results of component rotations, and selected the rotation method that results in more meaningful naming of the components based on the instrument.

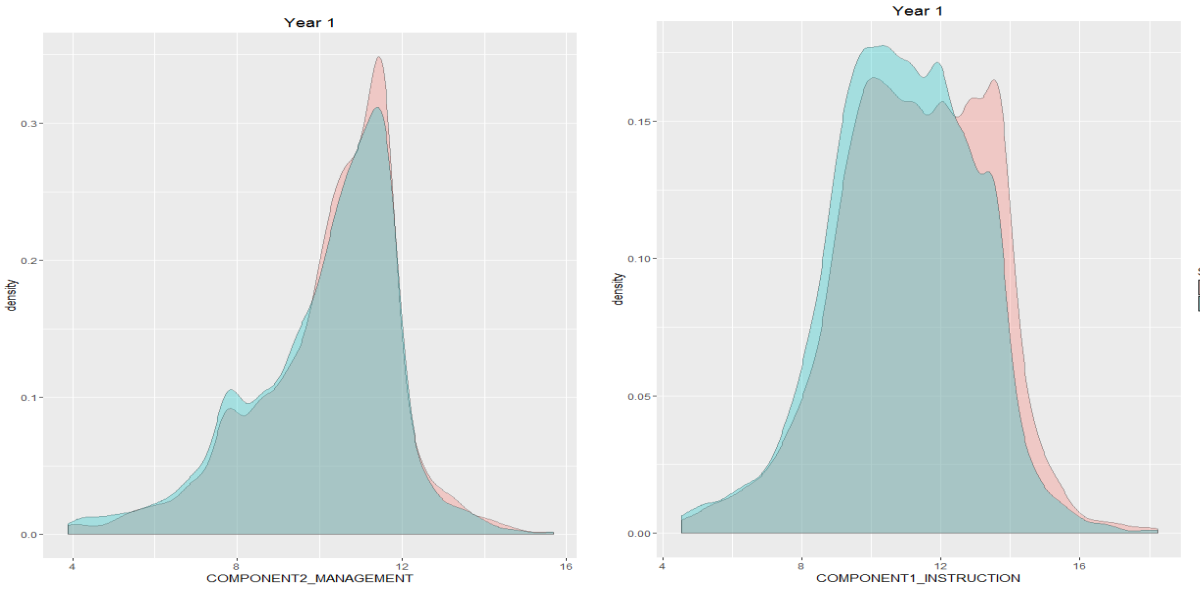
B.4. Interpretations of the PCA Results

For each instruments, the results of PCA is one or more components to represent different aspects of teacher quality that the instrument measures. Each component is a linear combination of all the dimensions (variables) in the instrument, with different coefficients as loadings. In other words, each component is now a composite measure that can be used to compute component scores. The component scores are calculated for each lesson on each component as representations of the corresponding teacher's quality in certain aspects, based on the meaning of the component. The same teacher's quality as captured and represented by different instruments/different content can now be compared and contrast using the component scores.

APPENDIX C:

SCORE DISTRIBUTION WITH THE FULL SAMPLE

Figure A 1: FFT raw component and simple average scores distribution: Year One



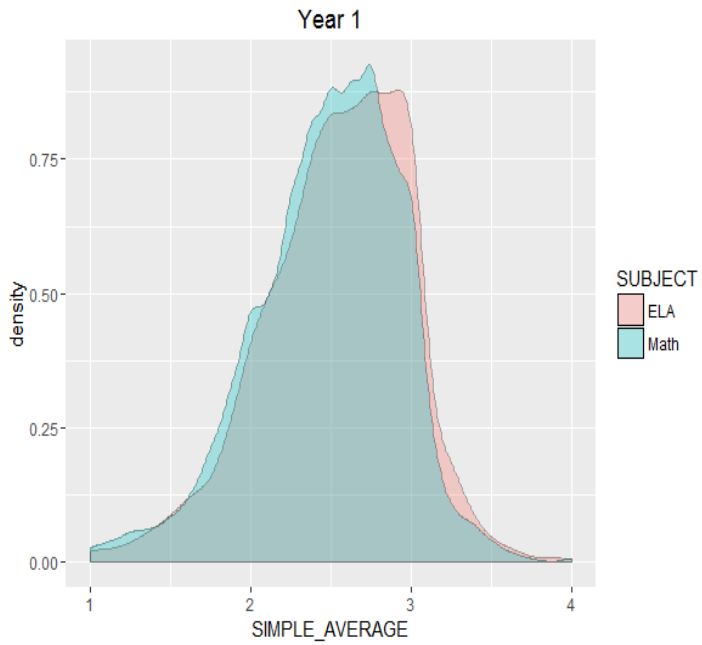


Figure A 2: FFT raw component and simple average scores distribution: Year Two

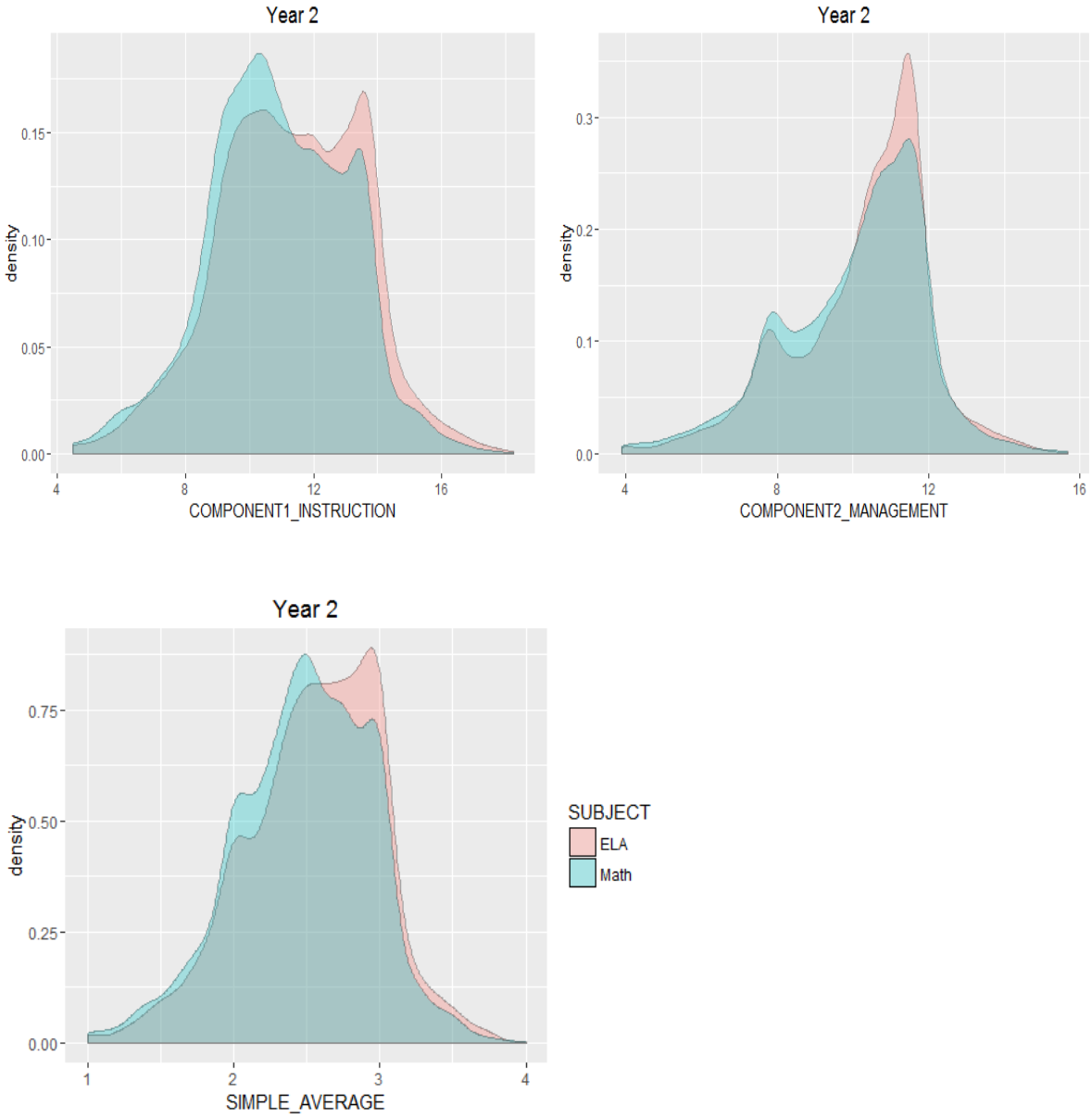


Figure A 3: CLASS raw component and simple average scores distribution: Year One

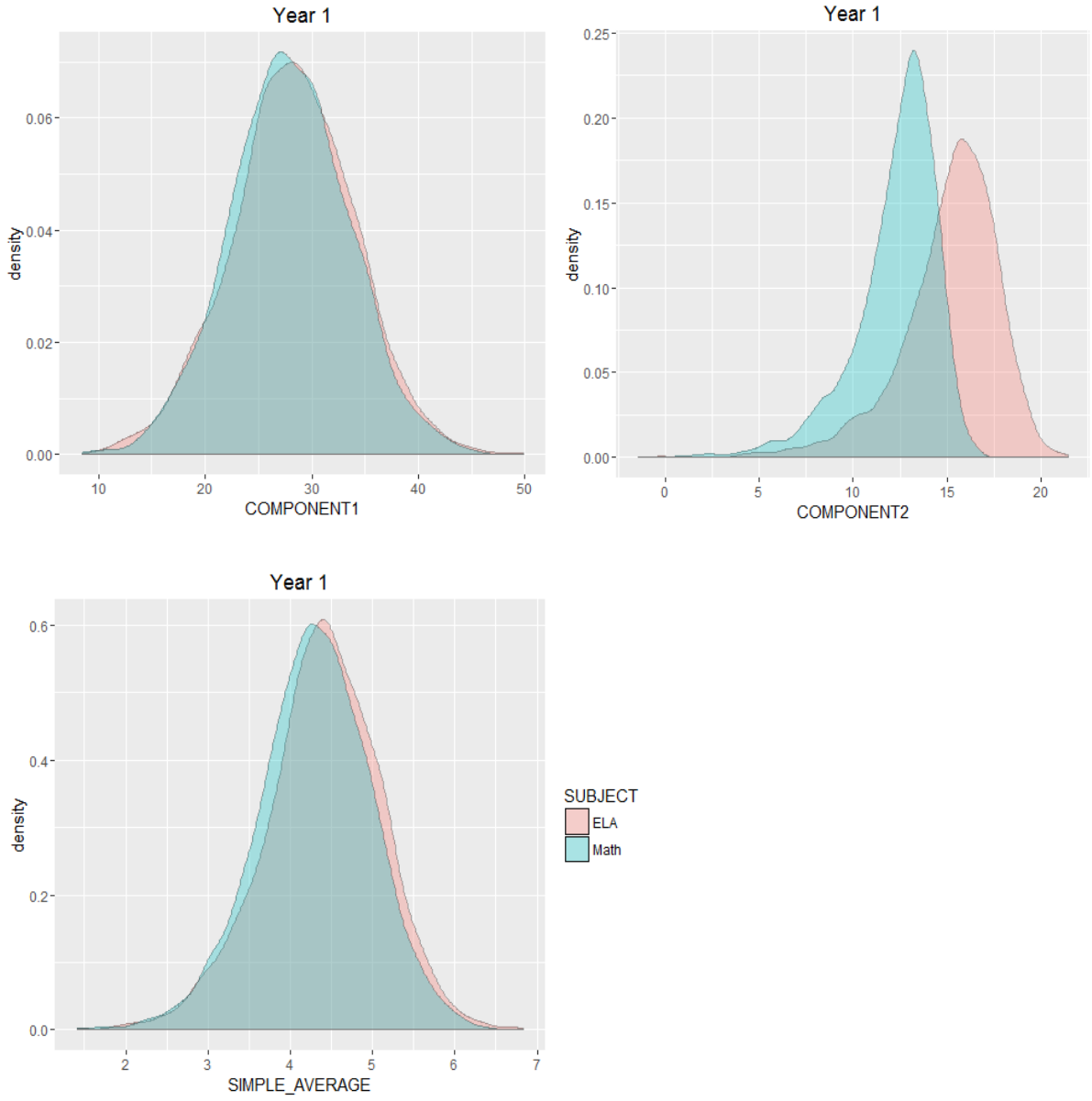


Figure A 4: CLASS raw component and simple average scores distribution: Year Two

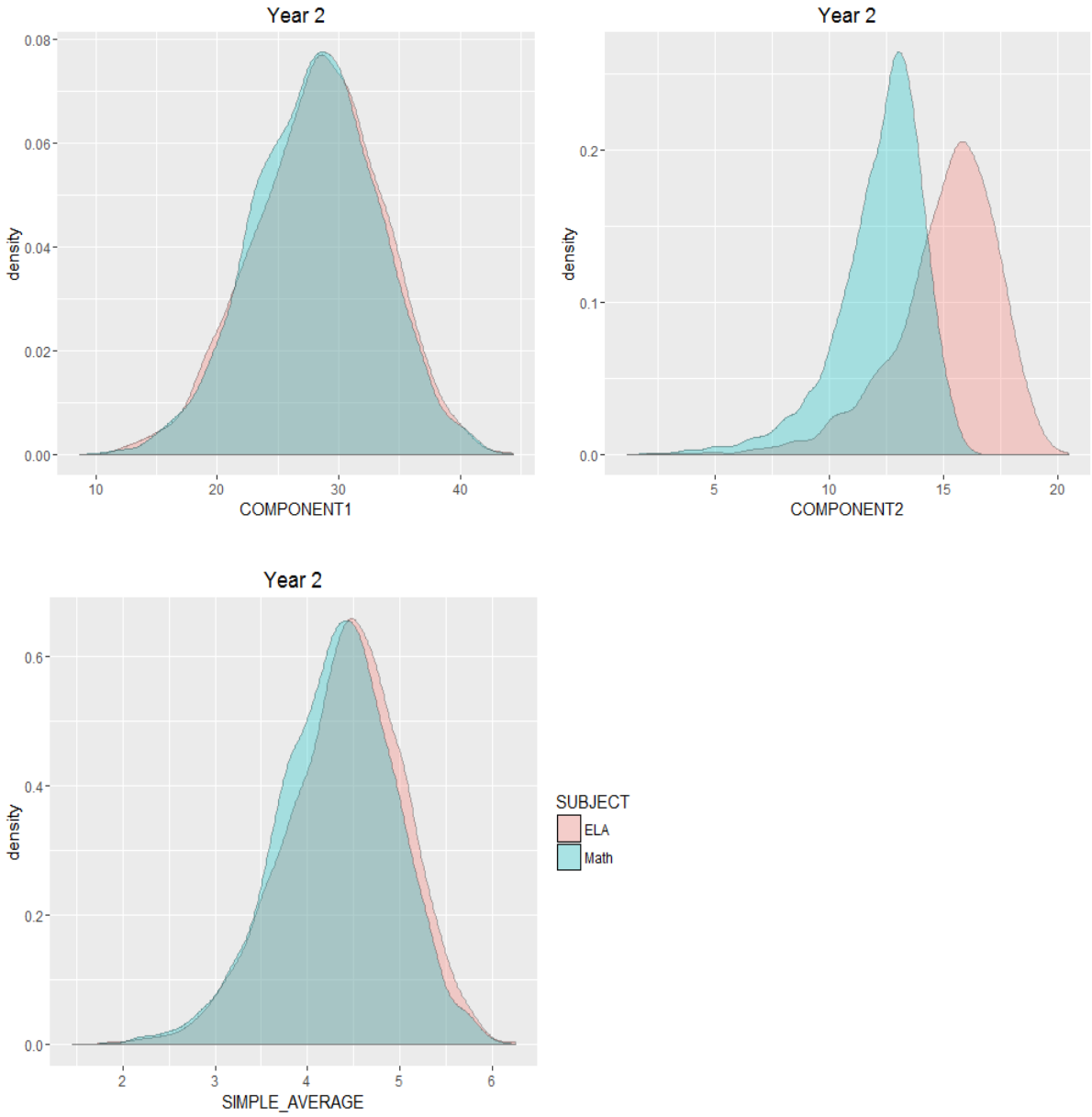
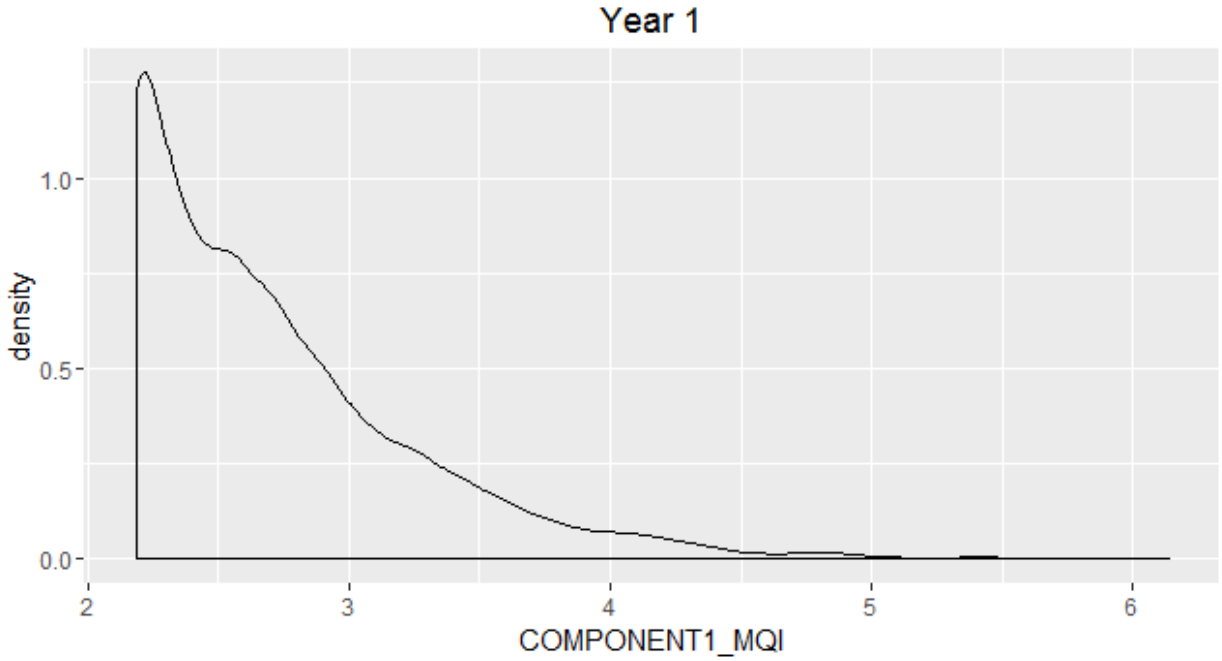


Figure A 5: MQI raw component scores distribution: Year One



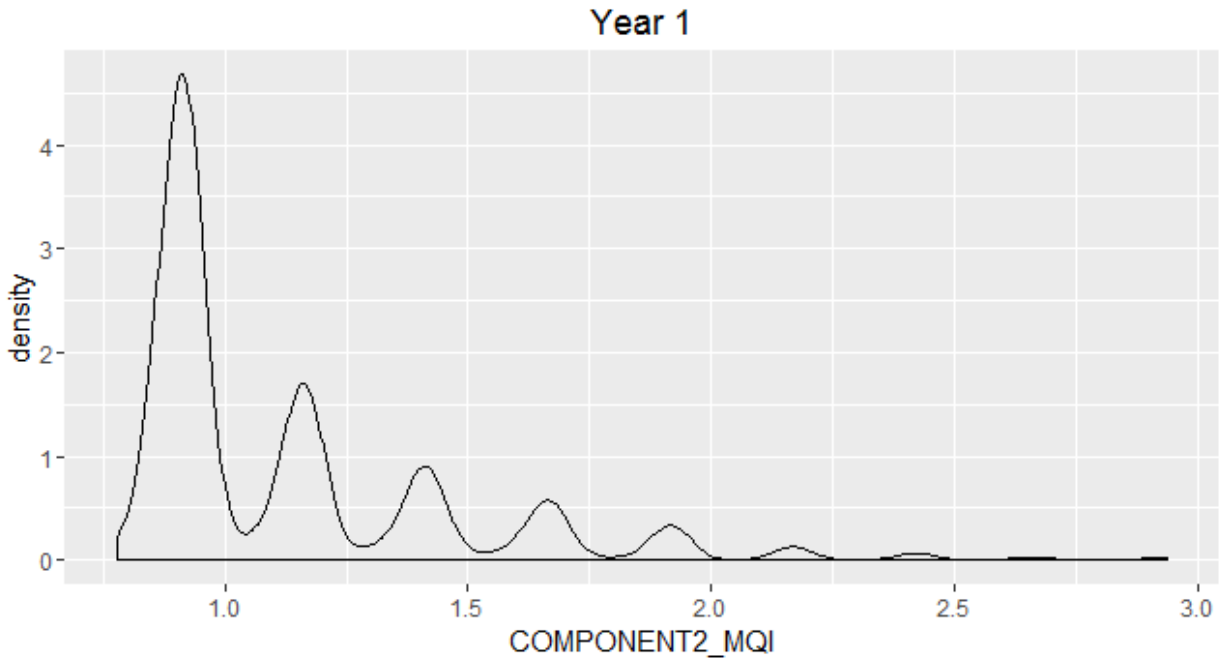


Figure A 6: MQI simple average scores distribution: Year One

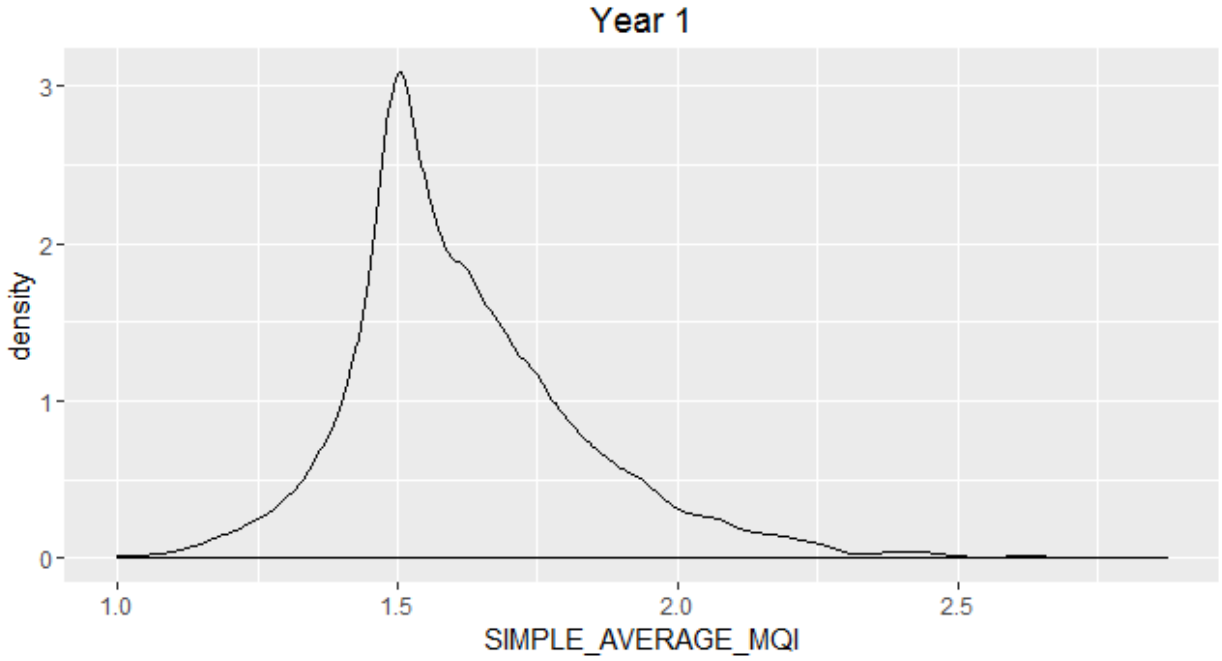


Figure A 7: MQI raw component scores distribution: Year Two

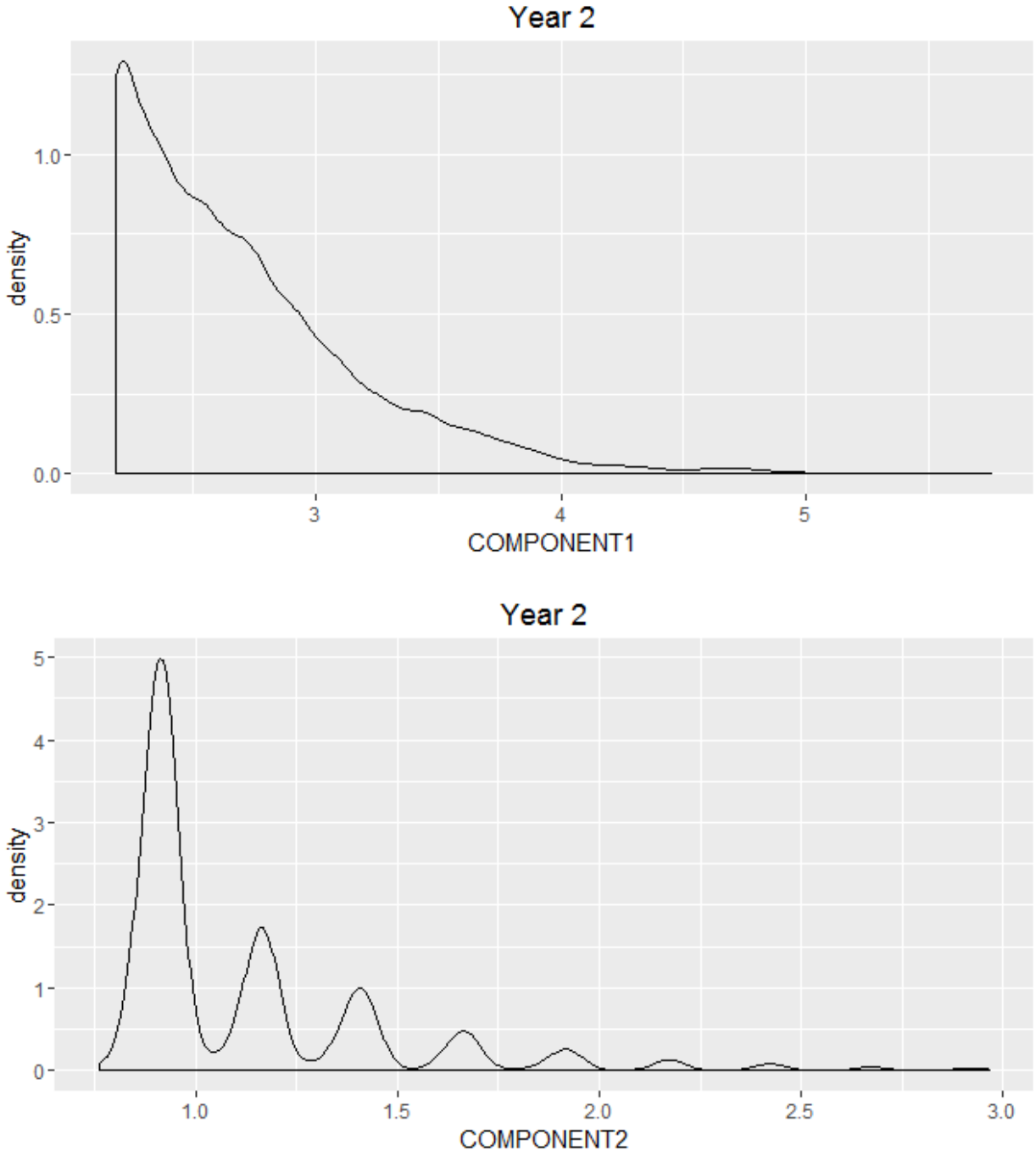


Figure A 8: MQI simple average scores distribution: Year Two

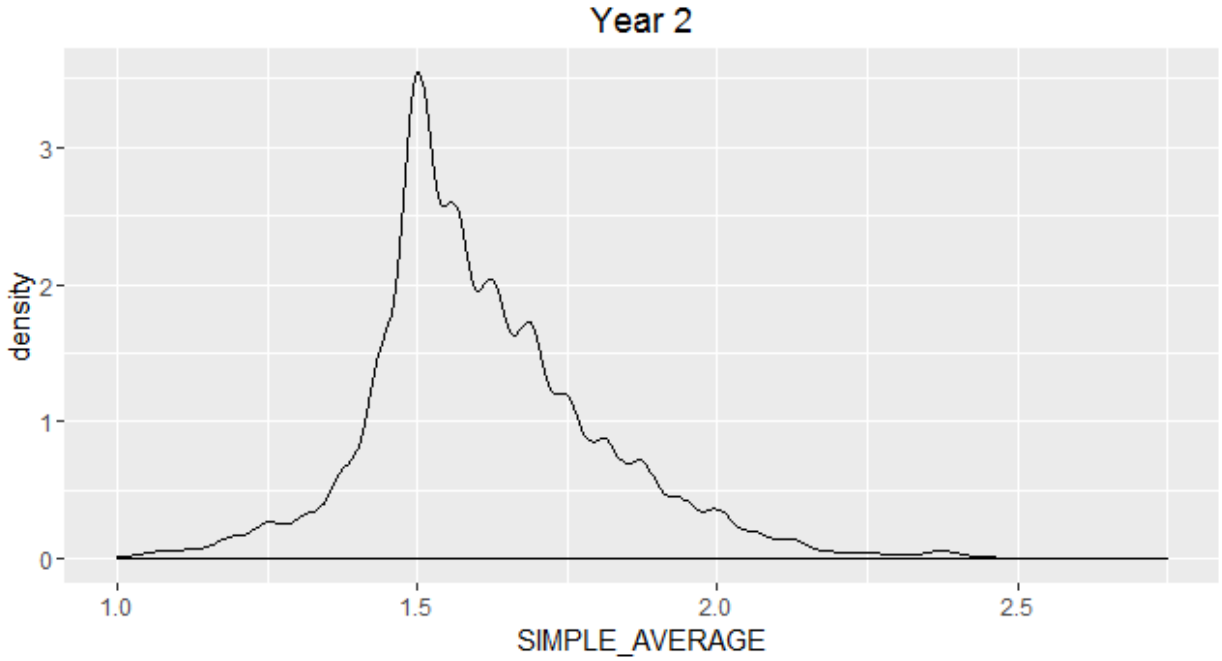
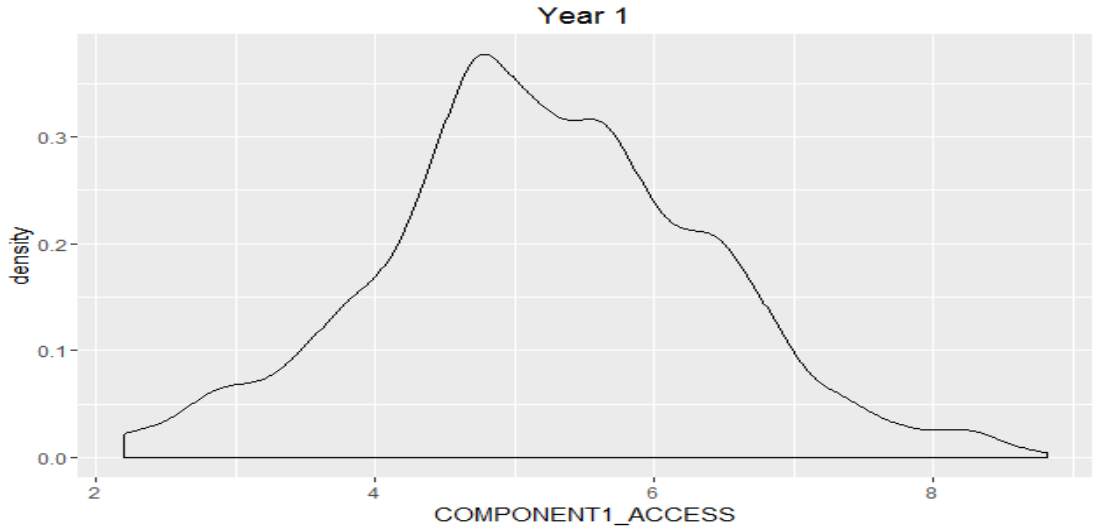
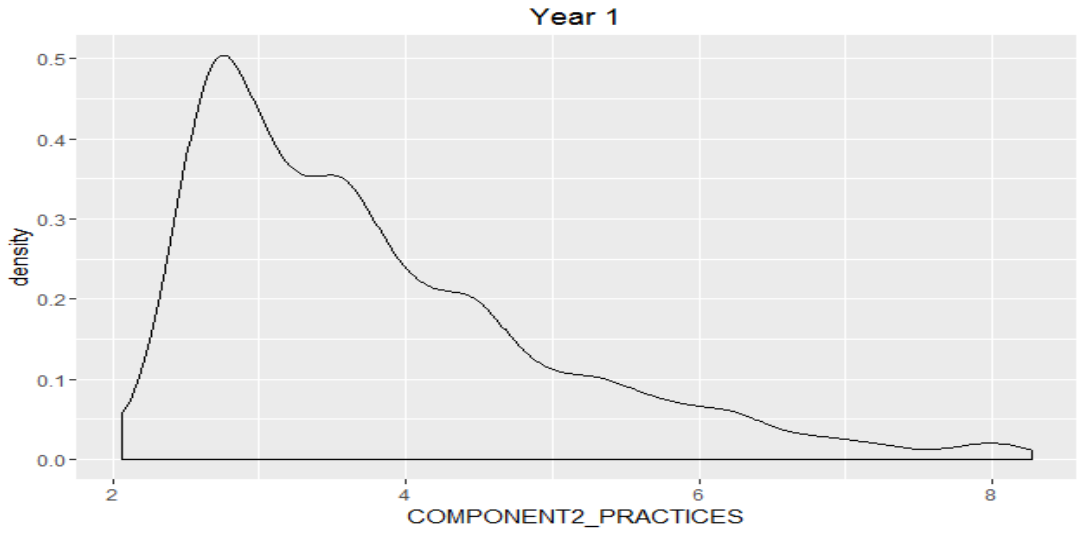


Figure A 9: PLATO raw component scores distribution: Year One





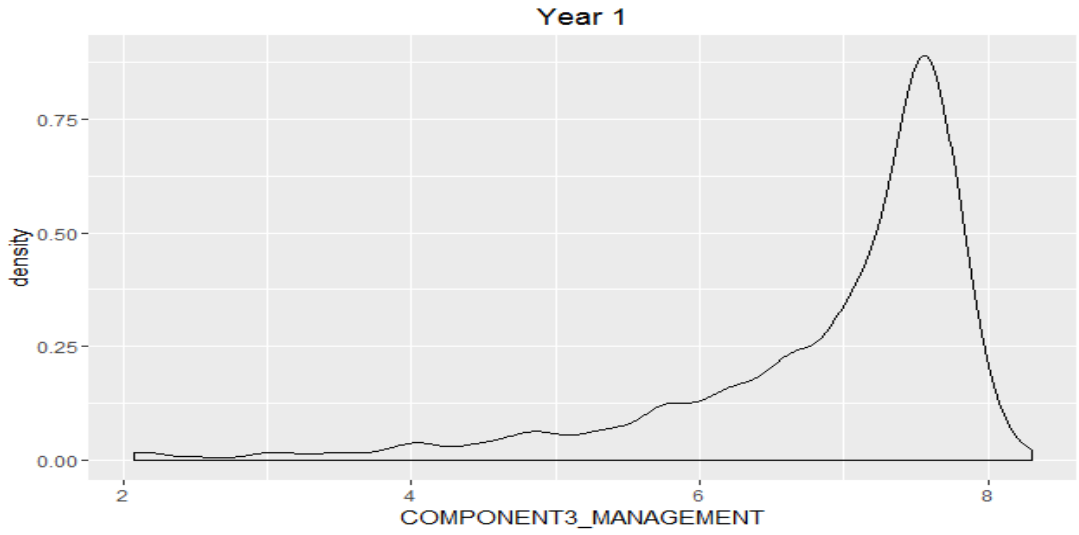


Figure A 10: PLATO simple average scores distribution: Year One

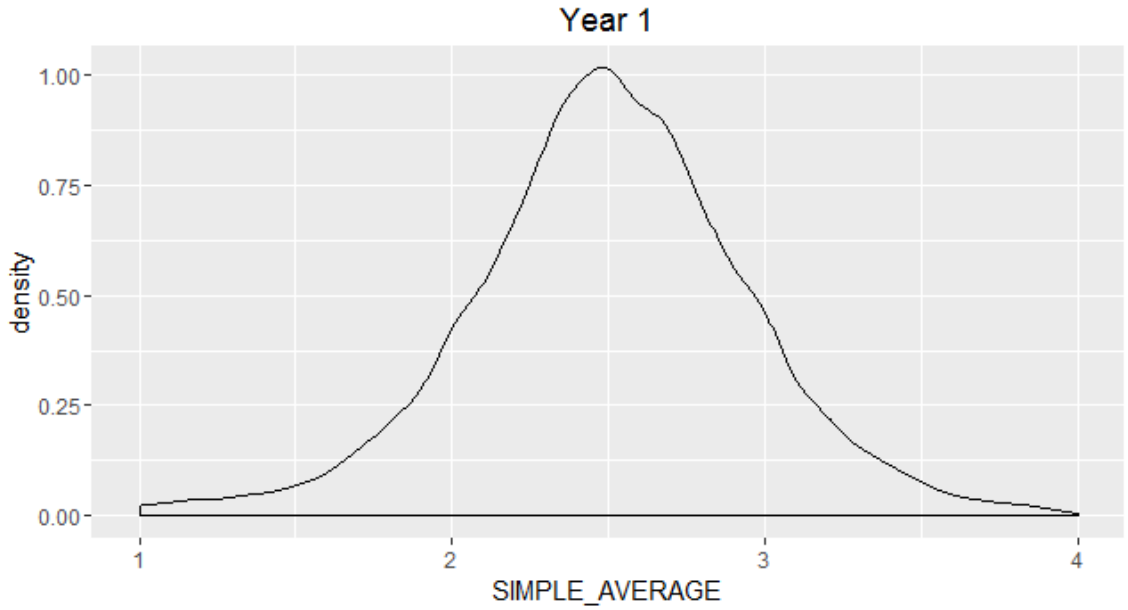


Figure A 11: PLATO raw component scores distribution: Year Two

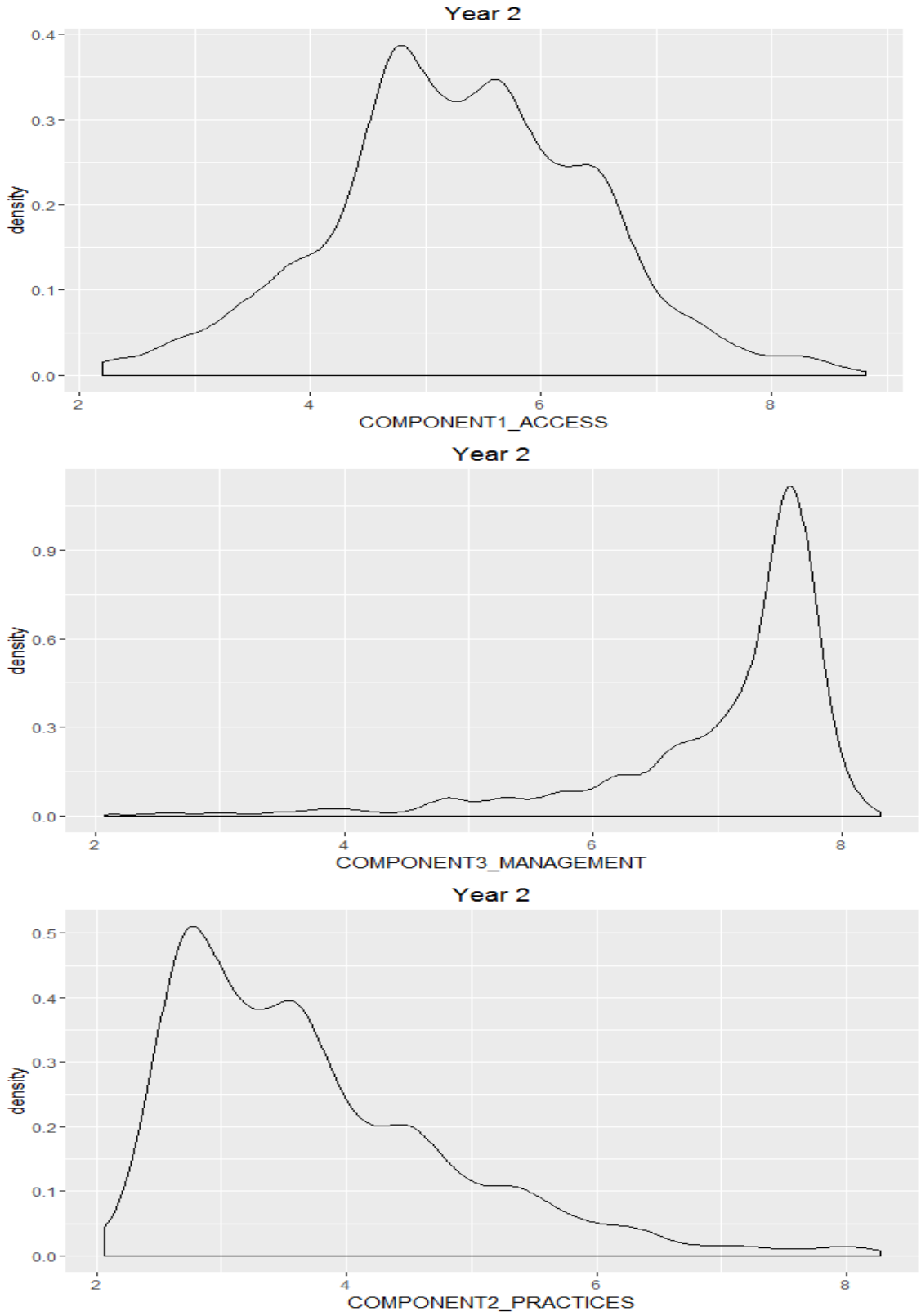
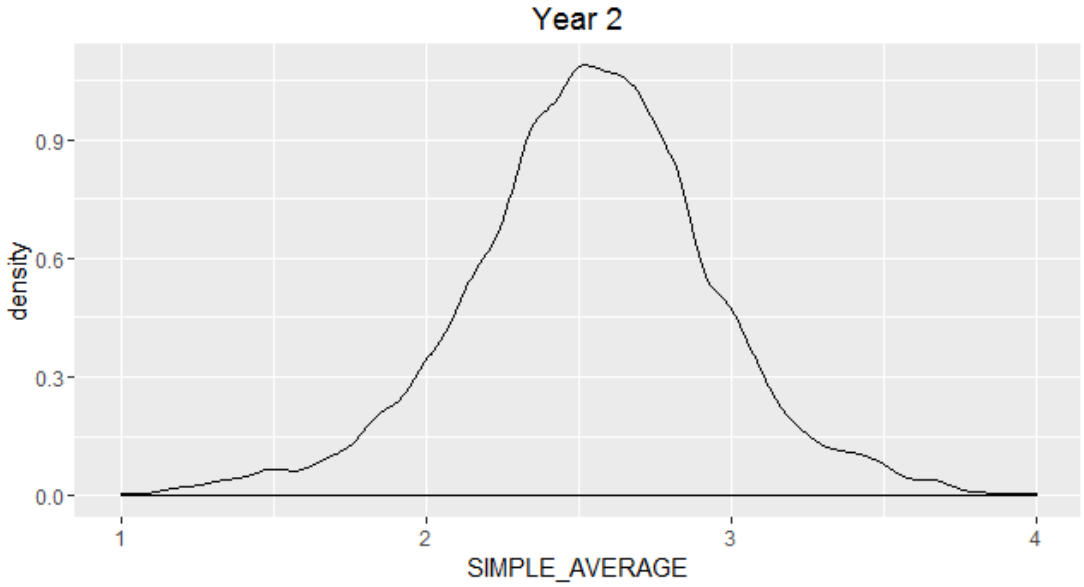


Figure A 12: PLATO simple average score distribution: Year Two



APPENDIX D:

COMPLETE LISTS OF COMPARISONS FOR EACH RESEARCH QUESTION

In this appendix, I provide a description of each pair of comparison under different contexts that I make to address the three research questions.

Research Question One: For the generalist teachers who taught both ELA and mathematics at elementary grades, to what extent are their observational scores different as measured by various protocols in the MET data? The comparisons I make include:

1. For each generalist teacher, compare his or her raw scores on FFT between mathematics and ELA:
    - Compare teachers' scores on FFT's first principal component ***Instruction*** between mathematics and ELA
    - Compare teachers' scores on FFT's second principal component ***Management*** between mathematics and ELA
    - Compare teachers' simple average scores across all FFT dimensions between mathematics and ELA
2. For each generalist teacher, compare his or her raw scores on CLASS between mathematics and ELA:
    - Compare teachers' scores on CLASS' first principal component ***Support*** between mathematics and ELA
    - Compare teachers' scores on CLASS' second principal component ***Organization*** between mathematics and ELA
    - Compare teachers' simple average scores across all CLASS dimensions between mathematics and ELA

3. For each generalist teacher, compare ranks (across all generalist teachers) on MQI for mathematics lessons to ranks on PLATO for ELA lessons:
    - Compare teachers' mathematics rank scores on MQI's first principal component of ***Instruction*** to ELA rank scores on PLATO's first principal component of ***Access***, which are one of the two pairs of conceptually matching components from MQI and PLATO, as identified in Chapter 4, section 4*;*
    - Compare teachers' mathematics rank scores on MQI's first principal component of ***Instruction*** to ELA rank scores on PLATO's second principal component of ***Practices*** which are the second pair of conceptually matching components from MQI and PLATO*;*
    - Compare teachers' simple average rank scores of mathematics across all MQI dimensions to simple average rank scores of ELA across all PLATO dimensions.

Research Question Two: For mathematics teachers, to what extent are their teacher quality measures different, as assessed by various observational instruments in the MET data? The comparison pairs I make include:

1. For each mathematics teacher, compare his or her mathematics lessons' scores on FFT to scores on CLASS:
    - Compare teachers' scores on FFT ***Instruction*** to scores on CLASS ***Support****;*
    - Compare teachers' scores on FFT ***Management*** to scores on CLASS ***Organization***;
    - Compare teachers' simple average scores on FFT to simple average scores on CLASS;

2. For each mathematics teacher, compare his or her mathematics lessons' scores on FFT to scores on MQI:
    - Compare teachers' scores on FFT ***Instruction*** to scores on MQI ***Instruction****;*
    - Compare teachers' simple average scores across all FFT dimensions to simple average scores across all MQI dimensions;
3. For each mathematics teacher, compare his or her mathematics lessons' scores on CLASS to scores on MQI:
    - Compare teachers' scores on CLASS ***Support*** to scores on MQI ***Instruction****;*
    - Compare teachers' simple average scores across all CLASS dimensions to simple average scores across all MQI dimensions;

Research Question Three: For mathematics teachers, to what extent do their teacher quality measures differ across subject areas within mathematics as assessed by relevant observational instruments? The comparisons I make include:

1. For each mathematics teacher, compare his or her raw scores on FFT between two subject areas within mathematics:
    - Compare teachers' scores on FFT ***Instruction*** between Algebra & Algebraic Thinking (AA) and Numbers & Operations (NO);
    - Compare teachers' scores on FFT ***Management*** between Algebra & Algebraic Thinking and Numbers & Operations;
    - Compare teachers' simple average scores on FFT between Algebra & Algebraic Thinking and Numbers & Operations;
    - Compare teachers' scores on FFT ***Instruction*** between Numbers & Operations (NO) and Geometry (G);

- Compare teachers' scores on FFT ***Management*** between Numbers & Operations and Geometry;
- Compare teachers' simple average scores on FFT between Numbers & Operations and Geometry;
- Compare teachers' scores on FFT ***Instruction*** between Algebra & Algebraic Thinking (AA) and Statistics & Probability (SP);
- Compare teachers' scores on FFT ***Management*** between Algebra & Algebraic Thinking and Statistics & Probability;
- Compare teachers' simple average scores on FFT between Algebra & Algebraic Thinking and Statistics & Probability;

2. For each mathematics teacher, compare his or her raw scores on CLASS between mathematics and ELA:
    - Compare teachers' scores on CLASS ***Support*** between Algebra & Algebraic Thinking (AA) and Numbers & Operations (NO);
    - Compare teachers' scores on CLASS ***Organization*** between Algebra & Algebraic Thinking and Numbers & Operations;
    - Compare teachers' simple average scores on CLASS between Algebra & Algebraic Thinking and Numbers & Operations;
    - Compare teachers' scores on CLASS ***Support*** between Numbers & Operations (NO) and Geometry (G);
    - Compare teachers' scores on CLASS ***Organization*** between Numbers & Operations and Geometry;

- Compare teachers' simple average scores on FFT between Numbers & Operations and Geometry;
- Compare teachers' scores on CLASS ***Support*** between Algebra & Algebraic Thinking (AA) and Statistics & Probability (SP);
- Compare teachers' scores on CLASS ***Organization*** between Algebra & Algebraic Thinking and Statistics & Probability;
- Compare teachers' simple average scores on CLASS between Algebra & Algebraic Thinking and Statistics & Probability;

3. For each mathematics teacher, compare his or her raw scores on MQI between two subject areas within mathematics:
    - Compare teachers' scores on MQI ***Instruction*** between Algebra & Algebraic Thinking (AA) and Numbers & Operations (NO);
    - Compare teachers' scores on MQI ***Accuracy*** between Algebra & Algebraic Thinking and Numbers & Operations;
    - Compare teachers' simple average scores on all MQI dimensions between Algebra & Algebraic Thinking and Numbers & Operations;
    - Compare teachers' scores on MQI ***Instruction*** between Numbers & Operations (NO) and Geometry (G);
    - Compare teachers' scores on MQI ***Accuracy*** between Numbers & Operations and Geometry;
    - Compare teachers' simple average scores on all MQI dimensions between Numbers & Operations and Geometry;

- Compare teachers' scores on MQI ***Instruction*** between Algebra & Algebraic Thinking (AA) and Statistics & Probability (SP);
- Compare teachers' scores on MQI ***Accuracy*** between Algebra & Algebraic Thinking and Statistics & Probability;
- Compare teachers' simple average scores on all MQI dimensions between Algebra & Algebraic Thinking and Statistics & Probability;

APPENDIX E:

FIGURES AND CORRELATIONS FOR CROSS-INSTRUMENTAL COMPARISONS

E.1. Generalist Teachers' Observational Ratings Across Subject-specific Instruments

Figure A 13: Scatterplots for each comparison in Year One and Year Two for PLATO vs. MQI

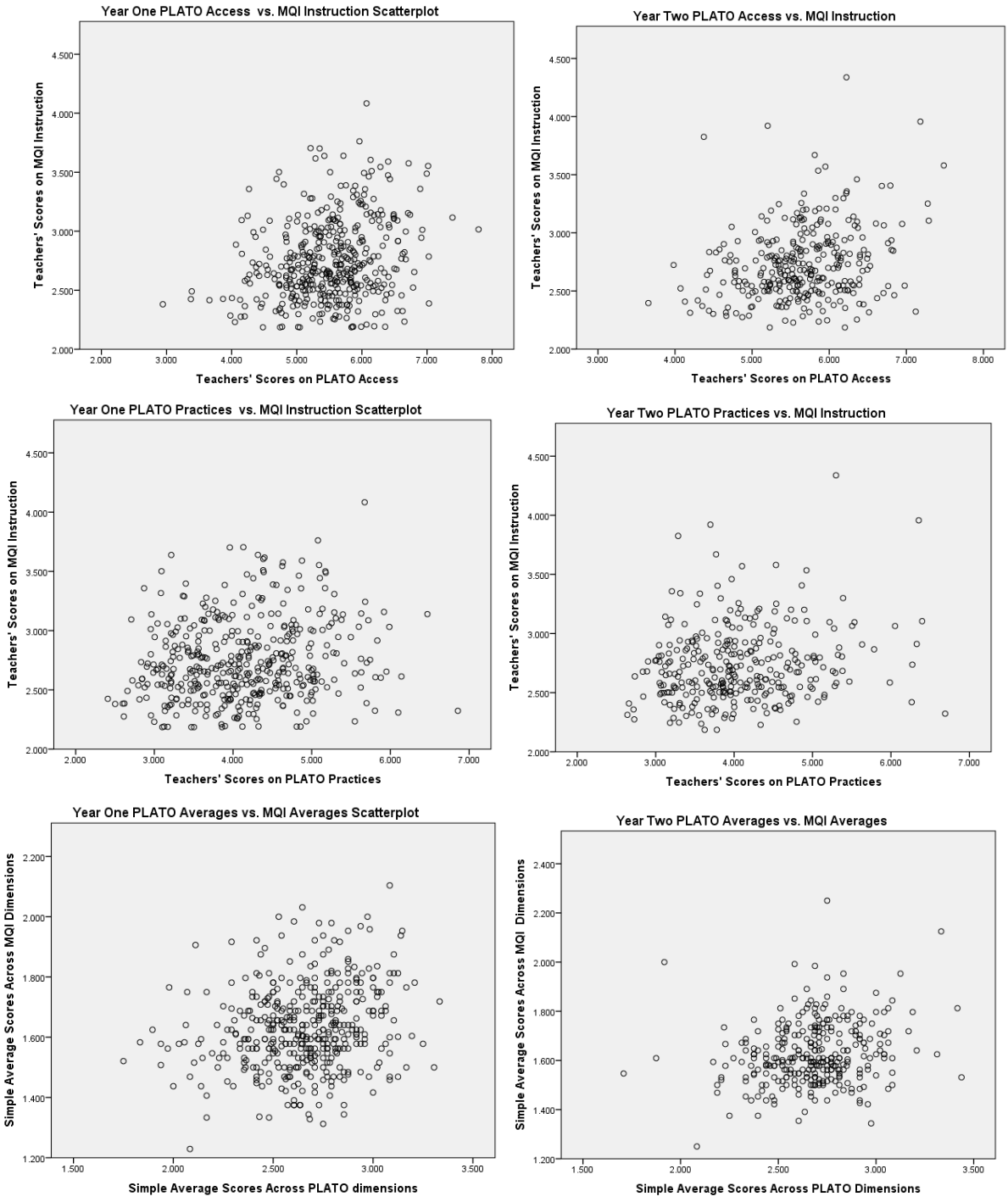


Table A 4: Correlation coefficient (Spearman's rho) from Spearman Rank Correlation tests

Instrument/ Content	Composite Score Generation Method	Comparison Level	Year One (N=430) p-value	Year Two (N=310) p-value
MQI math vs. PLATO ELA	PCA	MQI Instruction vs. PLATO Access	0.227**	0.224**
		MQI Instruction vs. PLATO Practices	0.161**	0.151**
	Simple average	MQI Average vs. PLATO Average	0.199***	0.196**

Note: ** means that correlation is significant at the 0.01 level (2-tailed). *** means that correlation is significant at the 0.001 level (2-tailed).

E.2. Comparisons of Mathematics Teachers' Observational Scores Across Instruments

Figure A 14: Scatterplots for each comparison in Year One and Year Two: FFT vs. CLASS

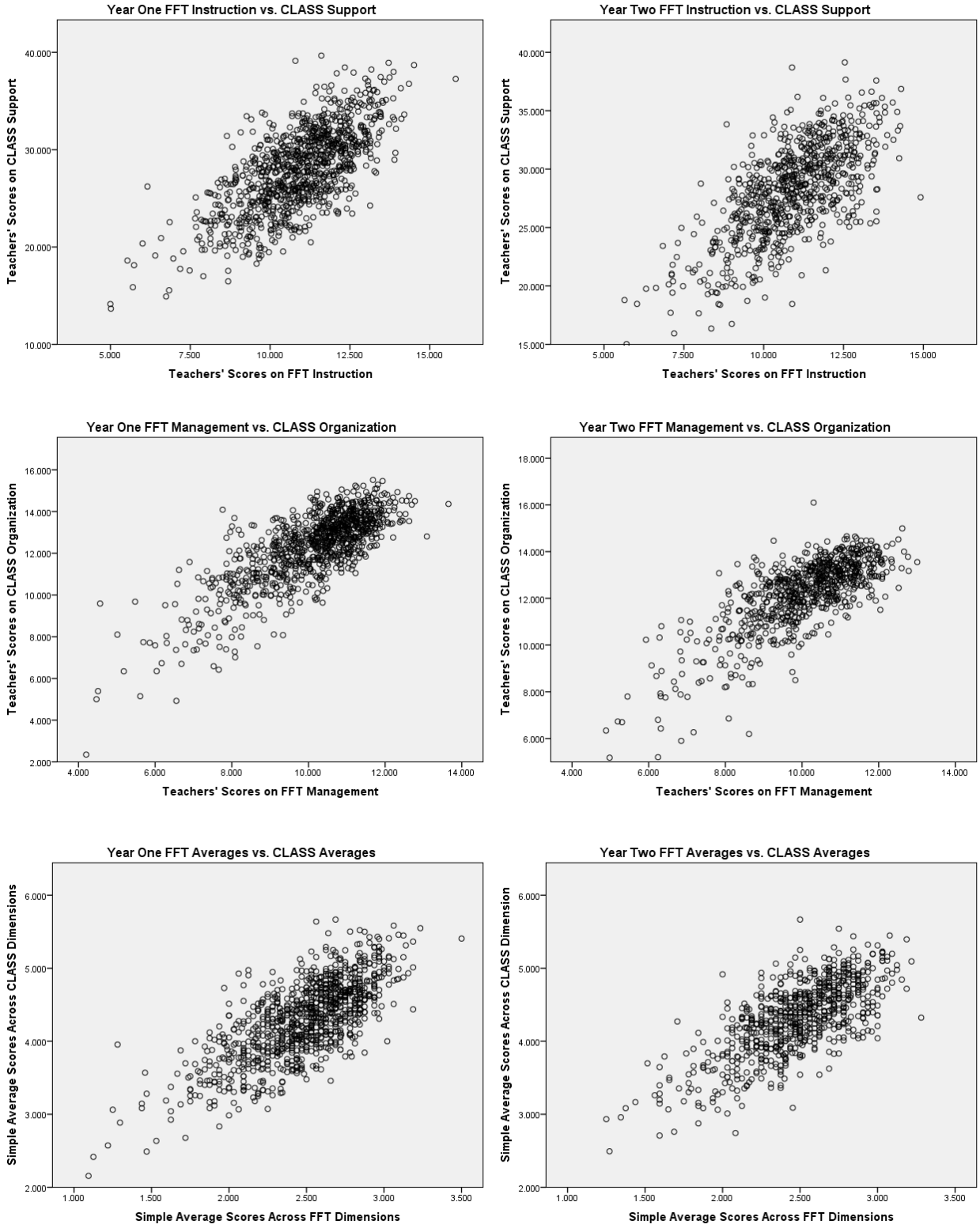


Figure A 15: Scatterplots for each pair of comparison in Year One and Year Two: FFT vs. MQI

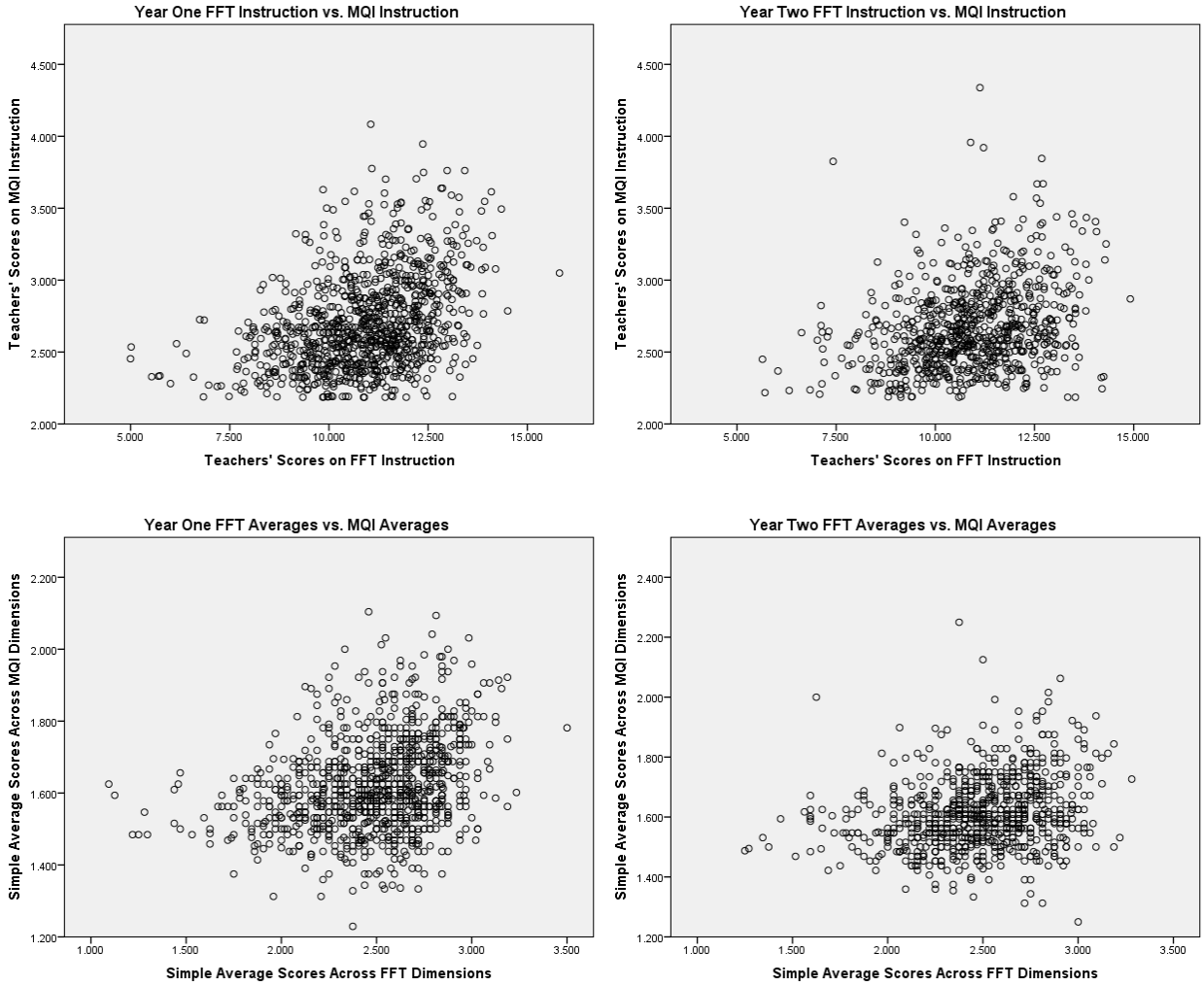


Figure A 16: Scatterplots for each comparison in Year One and Two: CLASS vs. MQI

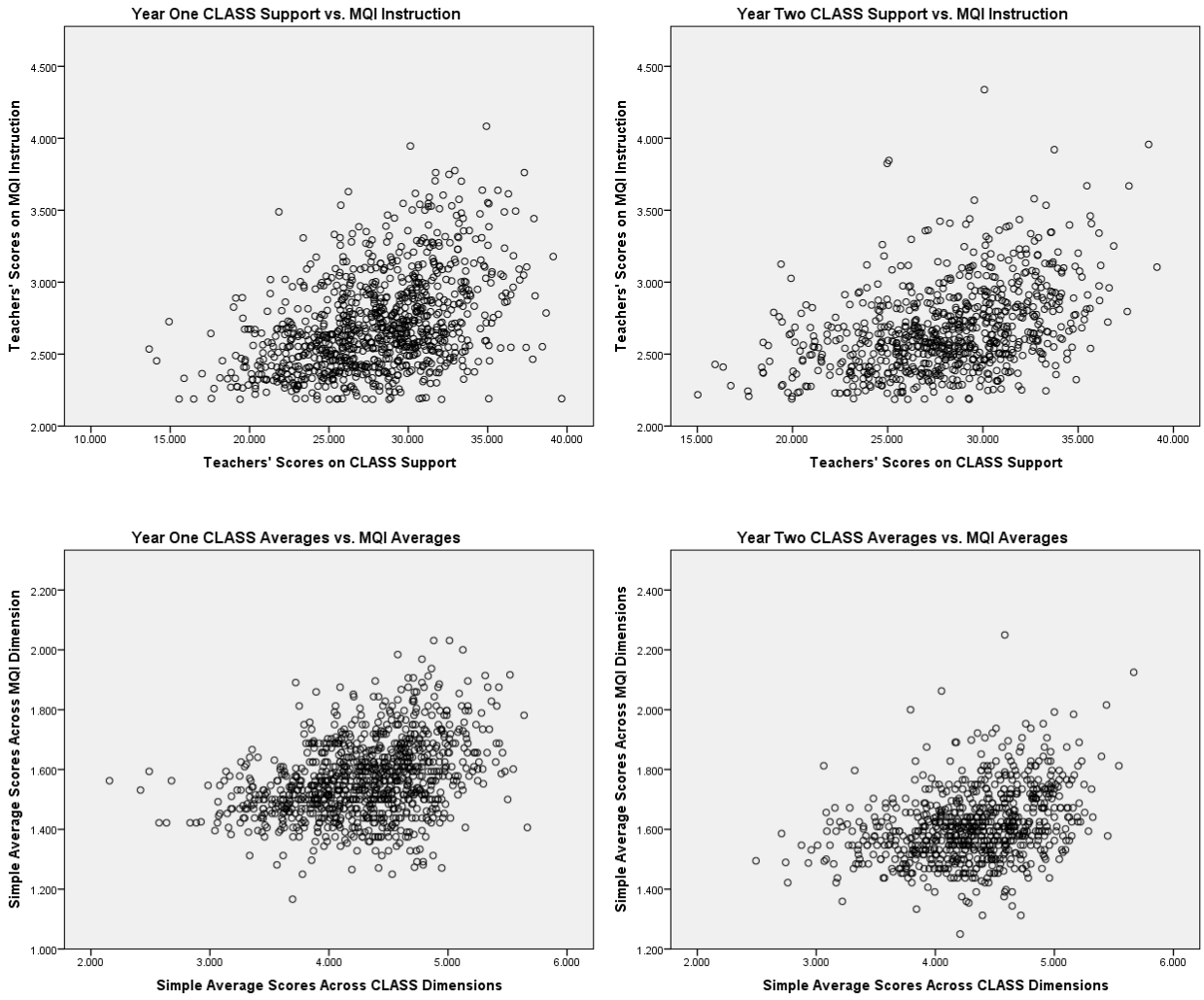


Table A 5: Correlation Coefficient (Spearman's rho) from Spearman Rank Correlation Tests

Instrument	Composite Score Generation Method	Comparison Level	Year One (N=978, 971) p-value	Year Two (N=770) p-value
FFT vs. CLASS	PCA	FFT Instruction vs. CLASS Support	0.668**	0.649**
		FFT Management vs. CLASS Organization	0.724**	0.701**
	Simple average	FFT Average vs. CLASS Average	0.694**	0.674**
MQI vs. Generic Instrument	PCA	FFT Instruction vs. MQI Instruction	0.374**	0.308**
		CLASS Support vs. MQI Instruction	0.420**	0.414**
	Simple Average	FFT Average vs. MQI Average	0.317**	0.259**
		CLASS Average vs. MQI Average	0.360**	0.339**

Note: ** means that correlation is significant at the 0.01 level (2-tailed).

APPENDIX F:

DIAGONAL ELEMENTS OF TRANSITION MATRICES FOR EACH COMPARISON

F.1. Generalist Teachers' Comparisons Across Subjects and Instruments

Table A 6: Diagonal elements in transition matrices for Year One generalist teachers: Generic instrument

Comp Level	Same Teacher's Decile Ranking Based on ELA Versus Math Lessons (N = 440)									
	1	2	3	4	5	6	7	8	9	10
FFT1	25.0%	15.9%	11.4%	13.6%	11.4%	13.6%	9.1%	13.6%	11.4%	29.5%
FFT2	43.2%	18.2%	13.6%	9.1%	11.4%	4.5%	13.6%	11.4%	4.5%	27.3%
FFT Ave	40.9%	12.8%	13.2%	13.3%	7.1%	20.8%	19.5%	11.8%	15.0%	27.3%
CLASS1	25.0%	22.7%	11.4%	9.1%	6.8%	9.1%	13.6%	11.4%	15.9%	43.2%
CLASS2	45.5%	11.4%	9.1%	13.6%	18.2%	13.6%	11.4%	11.4%	18.2%	40.9%
CLASS Ave	31.8%	11.4%	20.5%	22.7%	11.4%	13.6%	15.9%	11.6%	11.1%	38.6%

Note: ELA is the reference group.

Table A 7: Diagonal elements in transition matrices for Year Two generalist teachers: Generic instrument

Comp Level	Same Teacher's Decile Ranking Based on ELA Versus Math Lessons (N = 330)									
	1	2	3	4	5	6	7	8	9	10
FFT1	29.0%	9.7%	9.4%	12.9%	16.1%	12.5%	12.9%	21.9%	16.1%	32.3%
FFT2	32.3%	6.5%	12.5%	16.1%	3.2%	12.5%	6.5%	15.6%	16.1%	22.6%
FFT Ave	29.0%	16.1%	14.3%	5.6%	6.5%	9.1%	10.3%	21.2%	5.7%	26.9%
CLASS1	38.7%	12.9%	18.8%	19.4%	12.9%	15.6%	19.4%	15.6%	19.4%	29.0%
CLASS2	51.6%	9.7%	12.5%	6.5%	6.5%	9.4%	3.2%	12.5%	25.8%	22.6%
CLASS Ave	29.0%	13.8%	8.8%	19.4%	15.6%	9.7%	19.4%	15.6%	16.1%	32.3%

Note: ELA is the reference group.

Table A 8: Diagonal elements in transition matrices for generalist teachers: Subject-specific instrument

Decile	1	2	3	4	5	6	7	8	9	10
Year One (N=430)										
PLATO vs. MQI	18.6%	9.3%	18.6%	7.0%	9.3%	7.0%	11.6%	7.0%	20.9%	23.3%
	16.3%	11.6%	9.3%	11.6%	11.6%	11.6%	9.3%	18.6%	4.7%	16.3%
PLATO Ave vs. MQI Ave	15.9%	17.8%	10.3%	10.9%	20.5%	2.4%	5.4%	8.5%	18.6%	20.9%
Year Two (N=310)										
PLATO vs. MQI	25.8%	9.7%	16.1%	12.9%	12.9%	12.9%	6.5%	19.4%	16.1%	16.1%
	9.7%	12.9%	6.5%	16.1%	9.7%	9.7%	6.5%	6.5%	9.7%	9.7%
PLATO Ave vs. MQI Ave	15.6%	9.7%	11.1%	2.9%	10.8%	10.3%	0.0%	3.0%	13.3%	30.0%

Note: PLATO ***Access*** is compared to MQI ***Instruction*** in the first row of PLATO vs. MQI, and PLATO ***Practices*** is compared to MQI ***Instruction*** in the second row. PLATO is the reference group in each case.

F.2. Mathematics Teachers' Comparisons Across Instruments

Table A 9: Diagonal elements in transition matrices for Year One mathematics teachers

Sub	Total	FFT Instruction vs. CLASS Support									
		Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Math	978	45.4%	24.5%	21.4%	18.4%	15.3%	9.2%	12.4%	19.2%	19.4%	47.4%
FFT Management vs. CLASS Organization											
Math	978	63.9%	32.7%	19.4%	15.3%	17.3%	15.3%	18.4%	12.2%	16.3%	35.1%
FFT Average vs. CLASS Average											
Math	978	58.0%	27.2%	13.1%	16.0%	11.5%	14.1%	15.2%	19.4%	16.8%	47.4%
FFT Instruction vs. MQI Instruction											
Math	971	22.7%	13.4%	13.4%	11.3%	8.2%	11.2%	7.2%	10.3%	14.4%	26.8%
FFT Average vs. MQI Average											
Math	971	12.0%	10.3%	16.0%	8.6%	20.8%	10.1%	9.9%	12.6%	15.2%	28.9%
CLASS Support vs. MQI Instruction											
Math	971	28.9%	15.5%	17.5%	9.3%	9.3%	17.3%	11.3%	13.4%	16.5%	25.8%
CLASS Average vs. MQI Average											
Math	971	13.4%	11.3%	17.3%	6.3%	8.2%	9.2%	9.4%	18.6%	13.1%	31.3%

Note: The first instrument mentioned in the subtitles is the reference group for each pair.

Table A 10: Diagonal elements in transition matrices for Year Two mathematics teachers

Decile	FFT Instruction Versus CLASS Support (N = 772)									
	1	2	3	4	5	6	7	8	9	10
Math	59.7%	20.8%	15.6%	21.8%	14.3%	20.8%	19.2%	18.2%	19.5%	33.8%
FFT Management Versus CLASS Organization (N = 772)										
Math	61.0%	26.0%	18.2%	15.4%	15.6%	13.0%	14.1%	16.9%	16.9%	32.5%
FFT Average vs. CLASS Average (N = 772)										
Math	56.8%	15.6%	10.3%	14.8%	16.9%	16.5%	13.0%	15.0%	25.9%	38.6%
FFT Instruction Versus MQI Instruction (N = 770)										
Math	24.7%	15.6%	14.3%	9.1%	13.0%	5.3%	10.3%	7.8%	14.3%	19.5%
FFT Average vs. MQI Average (N = 770)										
Math	9.9%	13.0%	13.2%	9.1%	10.2%	11.2%	10.4%	13.8%	12.3%	21.4%
CLASS Support Versus MQI Instruction (N = 770)										
Math	24.7%	24.7%	13.0%	13.0%	15.6%	11.7%	14.3%	13.0%	23.4%	29.9%
CLASS Average vs. MQI Average (N = 770)										
Math	10.5%	17.7%	9.2%	9.0%	11.8%	6.7%	7.5%	8.0%	19.2%	24.7%

Note: The first instrument mentioned in the subtitles is the reference group for each pair.

F.3. Mathematics Teachers' Comparisons Across Subject Areas within Mathematics

Table A 11: Diagonal elements in transition matrices for Year One teachers who taught different subject areas within mathematics

	Same Teacher's Ranking Based on Algebra & Algebraic Thinking Versus Numbers & Operations Lessons (N = 230, 231, 221)									
Comp Level	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
FFT1	34.8%	12.5%	13.6%	11.1%	7.4%	16.7%	20.8%	10.0%	8.7%	12.0%
FFT2	41.7%	18.2%	8.7%	4.3%	21.7%	8.3%	4.3%	13.6%	12.5%	10.0%
FFT Ave	40.9%	20.0%	25.0%	11.1%	15.0%	9.7%	0.0%	16.7%	8.7%	11.1%
CLASS1	26.1%	26.1%	17.4%	13.0%	13.0%	4.2%	8.7%	17.4%	13.0%	30.4%
CLASS2	30.4%	21.7%	12.5%	9.1%	8.7%	16.7%	17.4%	13.0%	4.3%	21.7%
CLASS Ave	34.8%	27.3%	29.2%	8.0%	9.5%	4.2%	14.3%	16.0%	8.7%	26.1%
MQI1	20.0%	16.7%	21.7%	18.2%	18.2%	17.4%	9.1%	13.6%	4.5%	22.7%
MQI2	9.1%	18.2%	16.0%	15.8%	9.1%	12.5%	19.2%	0.0%	18.2%	18.2%
MQI Ave	15.0%	4.8%	5.3%	50.0%	8.0%	3.8%	15.8%	13.6%	13.6%	13.6%

Table A 11 (cont'd)

Same Teacher's Ranking Based on Numbers & Operations Versus Geometry Lessons (N = 135, 134, 125)										
Comp Level	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
	Quintile 1		Quintile 2		Quintile 3		Quintile 4		Quintile 5	
FFT1	7.7%	7.1%	7.7%	0.0%	15.4%	0.0%	15.4%	0.0%	22.2%	22.2%
FFT2	38.5%	7.1%	7.7%	6.7%	16.7%	20.0%	15.4%	23.1%	21.4%	15.4%
FFT Ave	37.5%		17.9%		27.3%		18.9%		41.7%	
CLASS1	38.5%	7.7%	7.1%	15.4%	0.0%	15.4%	0.0%	23.1%	7.1%	15.4%
CLASS2	30.8%	7.7%	7.1%	7.7%	7.1%	7.7%	14.3%	15.4%	7.1%	23.1%
CLASS Ave	30.8%	15.4%	13.3%	16.7%	0.0%	15.4%	0.0%	23.1%	7.1%	15.4%
MQI1	18.2%	7.1%	16.7%	0.0%	0.0%	33.3%	16.7%	15.4%	0.0%	0.0%
MQI2	8.3%	15.4%	16.7%	15.4%	16.7%	11.8%	25.0%	7.7%	15.4%	16.7%
MQI Ave	22.2%		28.1%		15.8%		22.9%		9.5%	

Note: For each groups of comparisons, Algebra & Algebraic Thinking and Geometry are the reference group respectively. FFT Average and MQI Average was divided into quintiles instead of deciles even though it has over 100 cases because there are too many ties. If divided into ten groups, there are multiple groups that have less than ten cases each, which is relatively uneven.

Table A 12: Diagonal elements in transition matrices for Year Two teachers who taught different subject areas within mathematics

Same Teacher's Ranking Based on Numbers & Operations Versus Algebra & Algebraic Thinking Lessons (N = 175, 175, 171)										
Comp Level	1	2	3	4	5	6	7	8	9	10
FFT1	64.7%	11.1%	15.8%	6.3%	5.9%	11.1%	16.7%	29.4%	16.7%	29.4%
FFT2	56.3%	10.5%	11.8%	11.1%	11.8%	16.7%	12.5%	15.8%	11.1%	35.3%
FFT Ave	47.1%	18.2%	14.3%	15.0%	6.3%	17.6%	25.0%	3.8%	10.5%	29.4%
CLASS1	52.9%	27.8%	5.9%	22.2%	17.6%	11.1%	5.6%	23.5%	27.8%	11.8%
CLASS2	35.3%	27.8%	23.5%	11.1%	11.8%	16.7%	22.2%	23.5%	5.9%	16.7%
CLASS Ave	52.9%	33.3%	11.1%	7.1%	15.0%	5.3%	12.5%	6.3%	23.8%	12.5%
MQI1	5.6%	6.3%	21.4%	15.0%	12.5%	5.3%	5.9%	23.5%	17.6%	11.8%
MQI2	5.9%	29.4%	17.6%	5.9%	5.9%	3.8%	11.1%	11.1%	6.3%	5.9%
MQI Average	6.7%	8.3%	12.0%	12.5%	22.2%	15.8%	0.0%	35.3%	13.3%	16.7%

Table A 12 (cont'd)

Same Teacher's Ranking Based on Numbers & Operations Versus Geometry Lessons (N = 84, 84, 81))					
Comp Level	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
FFT1	56.3%	11.8%	17.6%	41.2%	35.3%
FFT2	43.8%	11.8%	11.8%	23.5%	35.3%
FFT Ave	43.8%	19.0%	15.4%	33.3%	37.5%
CLASS1	50.0%	29.4%	17.6%	35.3%	23.5%
CLASS2	43.8%	17.6%	11.8%	23.5%	17.6%
CLASS Ave	43.8%	17.6%	17.6%	16.7%	6.3%
MQI1	12.5%	31.3%	17.6%	12.5%	18.8%
MQI2	12.5%	25.0%	17.6%	12.5%	18.8%
MQI Ave	0.0%	15.4%	31.3%	0.0%	6.7%

Note: For each groups of comparisons, Algebra & Algebraic Thinking and Geometry are the reference group respectively.

Table A 13: Diagonal elements in transition matrices for teachers who taught both algebra and statistics from Both Years' Sample

	Same Teacher's Ranking Based on Algebra & Algebraic Thinking Versus Statistics & Probability Lessons (N = 56, 56, 55)				
Comp Level	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
FFT1	36.4%	33.3%	25.0%	41.7%	45.5%
FFT2	27.3%	30.8%	18.2%	41.7%	45.5%
FFT Ave	25.0%	27.3%	18.2%	38.5%	44.4%
CLASS1	27.3%	25.0%	25.0%	8.3%	45.5%
CLASS2	18.2%	8.3%	8.3%	25.0%	0%
CLASS Ave	16.7%	30.0%	36.4%	25.0%	45.5%
MQI1	18.2%	18.2%	18.2%	18.2%	18.2%
MQI2	36.4%	27.3%	45.5%	25.0%	40.0%
MQI Ave	50%	23.1%	8.3%	10.0%	0.0%

Note: Algebra & Algebraic Thinking is the reference group.

APPENDIX G:

FREQUENCY TABLE FOR TEACHERS WHO REMAIN IN THE SAME PERCENTILE GROUP

G.1. Generalist Teachers' Ratings Across Subjects and Instruments

In the tables, Group 1 represents the bottom percentile group, Group 2 represents the second to last percentile group, and so on.

Table A 14: Frequency of teachers in each percentile group for FFT component scores comparisons: Year One

	FFT_Y1_ELA_Math (ELA as reference group)					
	Num in each group	ELA1_MATH1	% Unchanged	um in each grou	ELA2_MATH2	%Unchanged
Group 1	44	11	25.0%	44	19	43.2%
Group 2	44	7	15.9%	44	8	18.2%
Group 3	44	5	11.4%	44	6	13.6%
Group 4	44	6	13.6%	44	4	9.1%
Group 5	44	5	11.4%	44	5	11.4%
Group 6	44	6	13.6%	44	2	4.5%
Group 7	44	4	9.1%	44	6	13.6%
Group 8	44	6	13.6%	44	5	11.4%
Group 9	44	5	11.4%	44	2	4.5%
Group 10	44	13	29.5%	44	12	27.3%
Total	440	68	15.5%	440	69	15.7%

Table A 15: Frequency of teachers in each percentile group for FFT component scores comparisons: Year Two

	FFT_Y2_ELA_Math (ELA as reference group)					
	Num in each group	ELA1_MATH1	% Unchanged	lum in each grou	ELA2_MATH2	%Unchanged
Group 1	31	9	29.0%	31	10	32.3%
Group 2	31	3	9.7%	31	2	6.5%
Group 3	32	3	9.4%	32	4	12.5%
Group 4	31	4	12.9%	31	5	16.1%
Group 5	31	5	16.1%	31	1	3.2%
Group 6	32	4	12.5%	32	4	12.5%
Group 7	31	4	12.9%	31	2	6.5%
Group 8	32	7	21.9%	32	5	15.6%
Group 9	31	5	16.1%	31	5	16.1%
Group 10	31	10	32.3%	31	7	22.6%
Total	313	54	17.3%	313	45	14.4%

Table A 16: Frequency of teachers in each percentile group for FFT simple average scores comparisons: Year One and Year Two

	FFT_ELA_Math (ELA as reference group)					
	Num in each group	Y1_ELA_MATH	% Unchanged	Num in each group	Y2_ELA_MATH	%Unchanged
Group 1	44	18	40.9%	31	9	29.0%
Group 2	47	6	12.8%	31	5	16.1%
Group 3	38	5	13.2%	28	4	14.3%
Group 4	45	6	13.3%	36	2	5.6%
Group 5	42	3	7.1%	31	2	6.5%
Group 6	48	7	14.6%	33	3	9.1%
Group 7	41	8	19.5%	29	3	10.3%
Group 8	51	6	11.8%	33	7	21.2%
Group 9	40	6	15.0%	35	2	5.7%
Group 10	44	12	27.3%	26	7	26.9%
Total	440	77	17.5%	313	44	14.1%

Table A 17: Frequency of teachers in each percentile group for CLASS component scores comparisons: Year One

	CLASS_Y1_ELA_Math (ELA as reference group)					
	Num in each group	ELA1_MATH	% Unchanged	Num in each group	ELA2_MATH	%Unchanged
Group 1	44	11	25.0%	44	20	45.5%
Group 2	44	10	22.7%	44	5	11.4%
Group 3	44	5	11.4%	44	4	9.1%
Group 4	44	4	9.1%	44	6	13.6%
Group 5	44	3	6.8%	44	8	18.2%
Group 6	44	4	9.1%	44	6	13.6%
Group 7	44	6	13.6%	44	5	11.4%
Group 8	44	5	11.4%	44	5	11.4%
Group 9	44	7	15.9%	44	8	18.2%
Group 10	44	19	43.2%	44	18	40.9%
Total	440	74	16.8%	440	85	19.3%

Table A 18: Frequency of teachers in each percentile group for CLASS component scores comparisons: Year Two

	CLASS_Y2_ELA_Math (ELA as reference group)					
	Num in each group	ELA1_MATH	% Unchanged	Num in each group	ELA2_MATH	%Unchanged
Group 1	31	12	38.7%	31	16	51.6%
Group 2	31	4	12.9%	31	3	9.7%
Group 3	32	6	18.8%	32	4	12.5%
Group 4	31	6	19.4%	31	2	6.5%
Group 5	31	4	12.9%	31	2	6.5%
Group 6	32	5	15.6%	32	3	9.4%
Group 7	31	6	19.4%	31	1	3.2%
Group 8	32	5	15.6%	32	4	12.5%
Group 9	31	6	19.4%	31	8	25.8%
Group 10	31	9	29.0%	31	7	22.6%
Total	313	63	20.1%	313	50	16.0%

Table A 19: Frequency of teachers in each percentile group for CLASS simple average scores comparisons: Year One and Year Two

	CLASS_ELA_Math (ELA as reference group)					
	Num in each group	Y1_ELA_MATH	% Unchanged	Num in each group	Y2_ELA_MATH	%Unchanged
Group 1	44	14	31.8%	31	9	29.0%
Group 2	44	5	11.4%	29	4	13.8%
Group 3	44	9	20.5%	34	3	8.8%
Group 4	44	10	22.7%	31	6	19.4%
Group 5	44	5	11.4%	32	5	15.6%
Group 6	44	6	13.6%	31	3	9.7%
Group 7	44	7	15.9%	31	6	19.4%
Group 8	43	5	11.6%	32	5	15.6%
Group 9	45	5	11.1%	31	5	16.1%
Group 10	44	17	38.6%	31	10	32.3%
Total	440	83	18.9%	313	56	17.9%

Table A 20: Frequency of teachers in each percentile group for PLATO vs. MQI component scores comparisons: Year One

	PLATO_MQI_Y1_Math (PLATO as reference group)					
	Num in each group	PLATO1_MQI1	% Unchanged	Num in each group	PLATO2_MQI1	%Unchanged
Group 1	43	8	18.6%	43	7	16.3%
Group 2	43	4	9.3%	43	6	14.0%
Group 3	43	6	14.0%	43	5	11.6%
Group 4	43	4	9.3%	43	4	9.3%
Group 5	43	8	18.6%	43	4	9.3%
Group 6	43	4	9.3%	43	3	7.0%
Group 7	43	7	16.3%	43	3	7.0%
Group 8	43	6	14.0%	43	8	18.6%
Group 9	43	9	20.9%	43	3	7.0%
Group 10	43	7	16.3%	43	5	11.6%
Total	430	63	14.7%	430	48	11.2%

Table A 21: Frequency of teachers in each percentile group for PLATO vs. MQI component scores comparisons: Year Two

	PLATO_MQI_Y2_Math (PLATO as reference group)					
	Num in each group	PLATO1_MQI1	% Unchanged	Num in each grou	PLATO2_MQI1	%Unchanged
Group 1	31	8	25.8%	31	3	9.7%
Group 2	31	3	9.7%	31	4	12.9%
Group 3	31	5	16.1%	31	2	6.5%
Group 4	31	4	12.9%	31	5	16.1%
Group 5	31	4	12.9%	31	3	9.7%
Group 6	31	4	12.9%	31	3	9.7%
Group 7	31	2	6.5%	31	2	6.5%
Group 8	31	6	19.4%	31	2	6.5%
Group 9	31	5	16.1%	31	3	9.7%
Group 10	31	5	16.1%	31	3	9.7%
Total	310	46	14.8%	310	30	9.7%

Table A 22: Frequency of teachers in each percentile group for PLATO vs. MQI simple average scores comparisons: Year One and Year Two

	PLATO_MQI_Math (PLATO as reference group)					
	Num in each group	Y1_PLATO_MQI	% Unchanged	Num in each group	Y2_PLATO_MQI	%Unchanged
Group 1	44	7	15.9%	32	5	15.6%
Group 2	45	8	17.8%	31	3	9.7%
Group 3	39	4	10.3%	27	3	11.1%
Group 4	46	5	10.9%	35	1	2.9%
Group 5	44	9	20.5%	37	4	10.8%
Group 6	42	1	2.4%	29	3	10.3%
Group 7	37	2	5.4%	26	0	0.0%
Group 8	47	4	8.5%	33	1	3.0%
Group 9	43	8	18.6%	30	4	13.3%
Group 10	43	9	20.9%	30	9	30.0%
Total	430	57	13.3%	310	33	10.6%

G.2. Mathematics Teachers' Ratings Across Subjects and Instruments

G.2.1. Between Two Generic Instruments

Table A 23: Frequency of teachers in each percentile group for FFT vs. CLASS component scores comparisons: Year One

	FFT_CLASS_Y1_Math (FFT as reference group)					
	Num in each group	FFT1_CLASS1	% Unchanged	Num in each group	FFT2_CLASS2	%Unchanged
Group 1	97	44	45.4%	97	62	63.9%
Group 2	98	24	24.5%	98	32	32.7%
Group 3	98	21	21.4%	98	19	19.4%
Group 4	98	18	18.4%	98	15	15.3%
Group 5	98	15	15.3%	98	17	17.3%
Group 6	98	9	9.2%	98	15	15.3%
Group 7	97	12	12.4%	98	18	18.4%
Group 8	99	19	19.2%	98	12	12.2%
Group 9	98	19	19.4%	98	16	16.3%
Group 10	97	46	47.4%	97	34	35.1%
Total	978	227	23.2%	978	240	24.5%

Table A 24: Frequency of teachers in each percentile group for FFT vs. CLASS component scores comparisons: Year Two

	FFT_CLASS_Y2_Math (FFT as reference group)					
	Num in each group	FFT1_CLASS1	% Unchanged	Num in each group	FFT2_CLASS2	%Unchanged
Group 1	77	46	59.7%	77	47	61.0%
Group 2	77	16	20.8%	77	20	26.0%
Group 3	77	12	15.6%	77	14	18.2%
Group 4	78	17	21.8%	78	12	15.4%
Group 5	77	11	14.3%	77	12	15.6%
Group 6	77	16	20.8%	77	10	13.0%
Group 7	78	15	19.2%	78	11	14.1%
Group 8	77	14	18.2%	77	13	16.9%
Group 9	77	15	19.5%	77	13	16.9%
Group 10	77	26	33.8%	77	25	32.5%
Total	772	188	24.4%	772	177	22.9%

Table A 25: Frequency of teachers in each percentile group for FFT vs. CLASS simple average scores comparisons: Year One and Year Two

	FFT_CLASS_Math (FFT as reference group)					
	Num in each group	Y1_FFT_CLASS	% Unchanged	Num in each group	Y2_FFT_CLASS	%Unchanged
Group 1	100	58	58.0%	81	46	56.8%
Group 2	103	28	27.2%	77	12	15.6%
Group 3	84	11	13.1%	68	7	10.3%
Group 4	106	17	16.0%	88	13	14.8%
Group 5	113	13	11.5%	59	10	16.9%
Group 6	85	12	14.1%	91	15	16.5%
Group 7	92	14	15.2%	77	10	13.0%
Group 8	103	20	19.4%	80	12	15.0%
Group 9	95	16	16.8%	81	21	25.9%
Group 10	97	46	47.4%	70	27	38.6%
Total	978	235	24.0%	772	173	22.4%

G.2.2. Between Generic and Math-Specific Instrument

Table A 26: Frequency of teachers in each percentile group for FFT vs. MQI component scores comparisons: Year One

	FFT_MQI_Y1_Math (FFT as reference group)		
	Num in each group	FFT1_MQI1	% Unchanged
Group 1	97	22	22.7%
Group 2	97	13	13.4%
Group 3	97	13	13.4%
Group 4	97	11	11.3%
Group 5	97	8	8.2%
Group 6	98	11	11.2%
Group 7	97	7	7.2%
Group 8	97	10	10.3%
Group 9	97	14	14.4%
Group 10	97	26	26.8%
Total	971	135	13.9%

Table A 27: Frequency of teachers in each percentile group for FFT vs. MQI component scores comparisons: Year Two

	FFT_MQI_Y2_Math (FFT as reference group)		
	Num in each group	FFT1_MQI1	% Unchanged
Group 1	77	19	24.7%
Group 2	77	12	15.6%
Group 3	77	11	14.3%
Group 4	77	7	9.1%
Group 5	77	10	13.0%
Group 6	76	4	5.3%
Group 7	78	8	10.3%
Group 8	77	6	7.8%
Group 9	77	11	14.3%
Group 10	77	15	19.5%
Total	770	103	13.4%

Table A 28: Frequency of teachers in each percentile group for FFT vs. MQI component scores comparisons: Year One and Year Two

	FFT_MQI_Math (FFT as reference group)					
	Num in each group	Y1_FFT_MQI	% Unchanged	Num in each group	Y2_FFT_MQI	% Unchanged
Group 1	100	12	12.0%	81	8	9.9%
Group 2	87	9	10.3%	77	10	13.0%
Group 3	100	16	16.0%	68	9	13.2%
Group 4	105	9	8.6%	88	8	9.1%
Group 5	77	16	20.8%	59	6	10.2%
Group 6	119	12	10.1%	89	10	11.2%
Group 7	91	9	9.9%	77	8	10.4%
Group 8	103	13	12.6%	80	11	13.8%
Group 9	92	14	15.2%	81	10	12.3%
Group 10	97	28	28.9%	70	15	21.4%
Total	971	138	14.2%	770	95	12.3%

Table A 29: Frequency of teachers in each percentile group for CLASS vs. MQI component scores comparisons: Year One

	CLASS_MQI_Y1_Math (CLASS as reference group)		
	Num in each group	CLASS1_MQI1	% Unchanged
Group 1	97	28	28.9%
Group 2	97	15	15.5%
Group 3	97	17	17.5%
Group 4	97	9	9.3%
Group 5	97	9	9.3%
Group 6	98	17	17.3%
Group 7	97	11	11.3%
Group 8	97	13	13.4%
Group 9	97	16	16.5%
Group 10	97	25	25.8%
Total	971	160	16.5%

Table A 30: Frequency of teachers in each percentile group for CLASS vs. MQI component scores comparisons: Year Two

	CLASS_MQI_Y2_Math (CLASS as reference group)		
	Num in each group	CLASS1_MQI1	% Unchanged
Group 1	77	19	24.7%
Group 2	77	19	24.7%
Group 3	77	10	13.0%
Group 4	77	10	13.0%
Group 5	77	12	15.6%
Group 6	77	9	11.7%
Group 7	77	11	14.3%
Group 8	77	10	13.0%
Group 9	77	18	23.4%
Group 10	77	23	29.9%
Total	770	141	18.3%

Table A 31: Frequency of teachers in each percentile group for CLASS vs. MQI simple average scores comparisons: Year One and Year Two

	CLASS_MQI_Math (CLASS as reference group)					
	Num in each group	Y1_CLASS_MQI	% Unchanged	Num in each group	Y2_CLASS_MQI	% Unchanged
Group 1	97	13	13.4%	76	8	10.5%
Group 2	97	11	11.3%	79	14	17.7%
Group 3	98	17	17.3%	76	7	9.2%
Group 4	96	6	6.3%	78	7	9.0%
Group 5	97	8	8.2%	76	9	11.8%
Group 6	98	9	9.2%	75	5	6.7%
Group 7	96	9	9.4%	80	6	7.5%
Group 8	97	18	18.6%	75	6	8.0%
Group 9	99	13	13.1%	78	15	19.2%
Group 10	96	30	31.3%	77	19	24.7%
Total	971	134	13.8%	770	96	12.5%

G.3. Mathematics Teachers' Ratings Across Subject Areas and Instruments

G.3.1. FFT

G.3.1.1. Algebra & Algebraic Thinking (AA) vs. Numbers & Operations (NO)

Table A 32: Frequency of teachers in each percentile group for FFT component scores comparisons between AA and NO: Year One

	FFT_Y1_AA_NO (AA as reference group)					
	Num in each group	AA1_NO1	% Unchanged	Num in each group	AA2_NO2	%Unchanged
Group 1	23	8	34.8%	24	10	41.7%
Group 2	24	3	12.5%	22	4	18.2%
Group 3	22	3	13.6%	23	2	8.7%
Group 4	18	2	11.1%	23	1	4.3%
Group 5	27	2	7.4%	23	5	21.7%
Group 6	24	4	16.7%	24	2	8.3%
Group 7	24	5	20.8%	23	1	4.3%
Group 8	20	2	10.0%	22	3	13.6%
Group 9	23	2	8.7%	16	2	12.5%
Group 10	25	3	12.0%	30	3	10.0%
Total	230	34	14.8%	230	33	14.3%

Table A 33: Frequency of teachers in each percentile group for FFT component scores comparisons between AA and NO: Year Two

	FFT_Y2_AA_NO (AA as reference group)					
	Num in each group	AA1_NO1	% Unchanged	Num in each group	AA2_NO2	%Unchanged
Group 1	17	11	64.7%	16	9	56.3%
Group 2	18	2	11.1%	19	2	10.5%
Group 3	19	3	15.8%	17	2	11.8%
Group 4	16	1	6.3%	18	2	11.1%
Group 5	17	1	5.9%	17	2	11.8%
Group 6	18	2	11.1%	18	3	16.7%
Group 7	18	3	16.7%	16	2	12.5%
Group 8	17	5	29.4%	19	3	15.8%
Group 9	18	3	16.7%	18	2	11.1%
Group 10	17	5	29.4%	17	6	35.3%
Total	175	36	20.6%	175	33	18.9%

Table A 34: Frequency of teachers in each percentile group for FFT simple average scores comparisons between AA and NO: Year One and Year Two

	FFT_AA_NO (AA as reference group)					
	Y1 Num in each group	Year One	% Unchanged	Y2 Num in each group	Year Two	%Unchanged
Group 1	22	9	40.9%	17	8	47.1%
Group 2	20	4	20.0%	22	4	18.2%
Group 3	24	6	25.0%	14	2	14.3%
Group 4	27	3	11.1%	20	3	15.0%
Group 5	20	3	15.0%	16	1	6.3%
Group 6	31	3	9.7%	17	3	17.6%
Group 7	6	0	0.0%	20	5	25.0%
Group 8	30	5	16.7%	13	4	30.8%
Group 9	23	2	8.7%	19	2	10.5%
Group 10	27	3	11.1%	17	5	29.4%
Total	230	38	16.5%	175	37	21.1%

G.3.1.2. Numbers & Operations (NO) vs. Geometry (G)

Table A 35: Frequency of teachers in each percentile group for FFT component scores comparisons between NO and G: Year One

	FFT_Y1_NO_G (NO as reference group)					
	Num in each group	NO1_G1	% Unchanged	Num in each group	NO2_G2	%Unchanged
Group 1	13	1	7.7%	13	5	38.5%
Group 2	14	1	7.1%	14	1	7.1%
Group 3	13	1	7.7%	13	1	7.7%
Group 4	14	0	0.0%	15	1	6.7%
Group 5	13	2	15.4%	12	2	16.7%
Group 6	15	0	0.0%	15	3	20.0%
Group 7	13	2	15.4%	13	2	15.4%
Group 8	13	0	0.0%	13	3	23.1%
Group 9	18	4	22.2%	14	3	21.4%
Group 10	9	2	22.2%	13	2	15.4%
Total	135	13	9.6%	135	23	17.0%

Table A 36: Frequency of teachers in each percentile group for FFT component scores comparisons between NO and G: Year Two

	FFT_Y2_NO_G (NO as reference group)					
	Num in each group	NO1_G1	% Unchanged	Num in each group	NO2_G2	%Unchanged
Group 1	16	9	56.3%	16	7	43.8%
Group 2	17	2	11.8%	17	2	11.8%
Group 3	17	3	17.6%	17	2	11.8%
Group 4	17	7	41.2%	17	4	23.5%
Group 5	17	6	35.3%	17	6	35.3%
Total	84	27	32.1%	84	21	25.0%

Table A 37: Frequency of teachers in each percentile group for FFT simple average scores comparisons between NO and G: Year One and Year Two

	FFT_NO_G (NO as reference group)					
	Y1 Num in each group	Year One	% Unchanged	Y2 Num in each group	Year Two	%Unchanged
Group 1	24	9	37.5%	16	7	43.8%
Group 2	28	5	17.9%	21	4	19.0%
Group 3	22	6	27.3%	13	2	15.4%
Group 4	37	7	18.9%	18	6	33.3%
Group 5	24	10	41.7%	16	6	37.5%
Total	135	37	27.4%	84	25	29.8%

G.3.1.3. Algebra & Algebraic Thinking (AA) vs. Statistics & Probability (SP)

Table A 38: Frequency of teachers in each percentile group for FFT component scores comparisons between AA and SP: Year One and Year Two

	FFT_Y1_Y2_AA_SP (AA as reference group)					
	Num in each group	AA1_SP1	% Unchanged	Num in each group	AA2_SP2	%Unchanged
Group 1	11	4	36.4%	11	3	27.3%
Group 2	12	4	33.3%	13	4	30.8%
Group 3	12	3	25.0%	11	2	18.2%
Group 4	12	5	41.7%	12	5	41.7%
Group 5	11	5	45.5%	11	5	45.5%
Total	58	21	36.2%	58	19	32.8%

Table A 39: Frequency of teachers in each percentile group for FFT simple average scores comparisons between AA and SP: Year One and Year Two

	FFT_AA_SP (AA as reference group)		
	Num in each group	Both Years	% Unchanged
Group 1	12	3	25.0%
Group 2	11	3	27.3%
Group 3	11	2	18.2%
Group 4	13	5	38.5%
Group 5	9	4	44.4%
Total	56	17	30.4%

G.3.2. CLASS

G.3.2.1. Algebra & Algebraic Thinking (AA) vs. Numbers & Operations (NO)

Table A 40: Frequency of teachers in each percentile group for CLASS component scores comparisons between AA and NO: Year One

	CLASS_Y1_AA_NO (AA as reference group)					
	Num in each group	AA1_NO1	% Unchanged	Num in each group	AA2_NO2	%Unchanged
Group 1	23	7	30.4%	23	7	30.4%
Group 2	23	6	26.1%	23	6	26.1%
Group 3	23	5	21.7%	23	3	13.0%
Group 4	24	3	12.5%	24	3	12.5%
Group 5	23	3	13.0%	23	2	8.7%
Group 6	23	1	4.3%	23	4	17.4%
Group 7	23	3	13.0%	23	5	21.7%
Group 8	23	4	17.4%	23	3	13.0%
Group 9	23	3	13.0%	23	1	4.3%
Group 10	23	7	30.4%	23	5	21.7%
Total	231	42	18.2%	231	39	16.9%

Table A 41: Frequency of teachers in each percentile group for CLASS component scores comparisons between AA and NO: Year Two

	CLASS_Y2_AA_NO (AA as reference group)					
	Num in each group	AA1_NO1	% Unchanged	Num in each group	AA2_NO2	%Unchanged
Group 1	17	10	58.8%	17	5	29.4%
Group 2	18	4	22.2%	18	6	33.3%
Group 3	17	1	5.9%	17	3	17.6%
Group 4	18	2	11.1%	18	0	0.0%
Group 5	17	4	23.5%	17	2	11.8%
Group 6	18	2	11.1%	18	2	11.1%
Group 7	18	1	5.6%	18	3	16.7%
Group 8	17	4	23.5%	17	4	23.5%
Group 9	18	5	27.8%	17	2	11.8%
Group 10	17	2	11.8%	18	3	16.7%
Total	175	35	20.0%	175	30	17.1%

Table A 42: Frequency of teachers in each percentile group for CLASS simple average scores comparisons between AA and NO: Year One and Year Two

	CLASS_AA_NO (AA as reference group)					
	Y1 Num in each group	Year One	% Unchanged	Y2 Num in each group	Year Two	%Unchanged
Group 1	23	8	34.8%	17	9	52.9%
Group 2	22	6	27.3%	18	6	33.3%
Group 3	24	7	29.2%	18	2	11.1%
Group 4	25	2	8.0%	14	1	7.1%
Group 5	21	2	9.5%	20	3	15.0%
Group 6	24	1	4.2%	19	1	5.3%
Group 7	21	3	14.3%	16	2	12.5%
Group 8	25	4	16.0%	16	1	6.3%
Group 9	23	2	8.7%	21	5	23.8%
Group 10	23	6	26.1%	16	2	12.5%
Total	231	41	17.7%	175	32	18.3%

G.3.2.2. Numbers & Operations (NO) vs. Geometry (G)

Table A 43: Frequency of teachers in each percentile group for CLASS component scores comparisons between NO and G: Year One

	CLASS_Y1_NO_G (NO as reference group)					
	Num in each group	NO1_G1	% Unchanged	Num in each group	NO2_G2	%Unchanged
Group 1	13	5	38.5%	13	4	30.8%
Group 2	13	1	7.7%	13	1	7.7%
Group 3	14	1	7.1%	14	2	14.3%
Group 4	13	2	15.4%	13	1	7.7%
Group 5	14	0	0.0%	14	0	0.0%
Group 6	13	2	15.4%	13	3	23.1%
Group 7	14	0	0.0%	14	2	14.3%
Group 8	13	3	23.1%	13	2	15.4%
Group 9	14	1	7.1%	14	1	7.1%
Group 10	13	2	15.4%	13	0	0.0%
Total	134	17	12.7%	134	16	11.9%

Table A 44: Frequency of teachers in each percentile group for CLASS component scores comparisons between NO and G: Year Two

	CLASS_Y2_NO_G (NO as reference group)					
	Num in each group	NO1_G1	% Unchanged	Num in each group	NO2_G2	%Unchanged
Group 1	16	8	50.0%	16	8	50.0%
Group 2	17	5	29.4%	17	3	17.6%
Group 3	17	3	17.6%	17	2	11.8%
Group 4	17	6	35.3%	17	5	29.4%
Group 5	17	3	17.6%	17	3	17.6%
Total	84	25	29.8%	84	21	25.0%

Table A 45: Frequency of teachers in each percentile group for CLASS simple average scores comparisons between NO and G: Year One and Year Two

	CLASS_NO_G (NO as reference group)					
	Y1 Num in each group	Year One	% Unchanged	Y2 Num in each group	Year Two	%Unchanged
Group 1	13	4	30.8%	16	7	43.8%
Group 2	13	2	15.4%	17	3	17.6%
Group 3	15	2	13.3%	17	3	17.6%
Group 4	12	2	16.7%	18	3	16.7%
Group 5	13	0	0.0%	16	1	6.3%
Group 6	13	2	15.4%			
Group 7	15	0	0.0%			
Group 8	13	3	23.1%			
Group 9	14	1	7.1%			
Group 10	13	2	15.4%			
Total	134	18	13.4%	84	17	20.2%

G.3.2.3. Algebra & Algebraic Thinking (AA) vs. Statistics & Probability (SP)

Table A 46: Frequency of teachers in each percentile group for CLASS component scores comparisons between AA and SP: Year One and Year Two

	CLASS_Y1_Y2_AA_SP (AA as reference group)					
	Num in each group	AA1_SP1	% Unchanged	Num in each group	AA2_SP2	%Unchanged
Group 1	11	3	27.3%	11	2	18.2%
Group 2	11	3	27.3%	11	1	9.1%
Group 3	12	4	33.3%	12	1	8.3%
Group 4	11	1	9.1%	11	3	27.3%
Group 5	11	5	45.5%	11	0	0.0%
Total	56	16	28.6%	56	7	12.5%

Table A 47: Frequency of teachers in each percentile group for CLASS simple average scores comparisons between AA and SP: Year One and Year Two

	CLASS_AA_SP (AA as reference group)		
	Num in each group	Both Years	% Unchanged
Group 1	12	2	16.7%
Group 2	10	3	30.0%
Group 3	11	4	36.4%
Group 4	12	3	25.0%
Group 5	11	5	45.5%
Total	56	17	30.4%

G.3.3. MQI

G.3.3.1. Algebra & Algebraic Thinking (AA) vs. Numbers & Operations (NO)

Table A 48: Frequency of teachers in each percentile group for MQI component scores comparisons between AA and NO: Year One

	MQI_Y1_AA_NO (AA as reference group)					
	Num in each group	AA1_NO1	% Unchanged	Num in each group	AA2_NO2	%Unchanged
Group 1	29	8	27.6%	22	2	9.1%
Group 2	16	1	6.3%	22	4	18.2%
Group 3	21	4	19.0%	23	4	17.4%
Group 4	21	5	23.8%	22	4	18.2%
Group 5	23	2	8.7%	21	3	14.3%
Group 6	23	3	13.0%	18	1	5.6%
Group 7	22	1	4.5%	29	8	27.6%
Group 8	22	2	9.1%	22	3	13.6%
Group 9	22	4	18.2%	20	3	15.0%
Group 10	22	4	18.2%	22	2	9.1%
Total	221	34	15.4%	221	34	15.4%

Table A 49: Frequency of teachers in each percentile group for MQI component scores comparisons between AA and NO: Year Two

	MQI_Y2_AA_NO (AA as reference group)					
	Num in each group	AA1_NO1	% Unchanged	Num in each group	AA2_NO2	%Unchanged
Group 1	18	1	5.6%	17	1	5.9%
Group 2	16	1	6.3%	17	5	29.4%
Group 3	14	3	21.4%	17	3	17.6%
Group 4	20	3	15.0%	17	1	5.9%
Group 5	16	2	12.5%	17	1	5.9%
Group 6	19	1	5.3%	26	1	3.8%
Group 7	17	1	5.9%	9	1	11.1%
Group 8	17	4	23.5%	18	2	11.1%
Group 9	17	3	17.6%	16	1	6.3%
Group 10	17	2	11.8%	17	1	5.9%
Total	171	21	12.3%	171	17	9.9%

Table A 50: Frequency of teachers in each percentile group for MQI simple average scores comparisons between AA and NO: Year One and Year Two

	MQI_AA_NO (AA as reference group)					
	Y1 Num in each group	Year One	% Unchanged	Y2 Num in each group	Year Two	%Unchanged
Group 1	20	3	15.0%	15	1	6.7%
Group 2	21	1	4.8%	12	1	8.3%
Group 3	38	2	5.3%	25	3	12.0%
Group 4	6	3	50.0%	8	1	12.5%
Group 5	25	2	8.0%	27	6	22.2%
Group 6	26	1	3.8%	19	3	15.8%
Group 7	19	3	15.8%	15	0	0.0%
Group 8	22	3	13.6%	17	6	35.3%
Group 9	22	3	13.6%	15	2	13.3%
Group 10	22	3	13.6%	18	3	16.7%
Total	221	24	10.9%	171	26	15.2%

G.3.3.2. Numbers & Operations (NO) vs. Geometry (G)

Table A 51: Frequency of teachers in each percentile group for MQI component scores comparisons between NO and G: Year One

	MQI_Y1_NO_G (NO as reference group)					
	Num in each group	NO1_G1	% Unchanged	Num in each group	NO2_G2	%Unchanged
Group 1	11	0	0.0%	12	1	8.3%
Group 2	13	2	15.4%	13	2	15.4%
Group 3	13	0	0.0%	12	3	25.0%
Group 4	13	0	0.0%	13	1	7.7%
Group 5	13	1	7.7%	12	2	16.7%
Group 6	12	2	16.7%	17	2	11.8%
Group 7	13	3	23.1%	9	3	33.3%
Group 8	12	2	16.7%	12	1	8.3%
Group 9	13	1	7.7%	13	2	15.4%
Group 10	12	1	8.3%	12	1	8.3%
Total	125	12	9.6%	125	18	14.4%

Table A 52: Frequency of teachers in each percentile group for MQI component scores comparisons between NO and G: Year Two

	MQI_Y2_NO_G (NO as reference group)					
	Num in each group	NO1_G1	% Unchanged	Num in each group	NO2_G2	%Unchanged
Group 1	16	2	12.5%	16	2	12.5%
Group 2	16	5	31.3%	16	4	25.0%
Group 3	17	3	17.6%	17	3	17.6%
Group 4	16	2	12.5%	16	2	12.5%
Group 5	16	3	18.8%	16	3	18.8%
Total	81	15	18.5%	81	14	17.3%

Table A 53: Frequency of teachers in each percentile group for FFT component scores comparisons between NO and G: Year One and Year Two

	MQI_NO_G (NO as reference group)					
	Y1 Num in each group	Year One	% Unchanged	Y2 Num in each group	Year Two	%Unchanged
Group 1	18	4	22.2%	19	0	0.0%
Group 2	32	9	28.1%	13	2	15.4%
Group 3	19	3	15.8%	16	5	31.3%
Group 4	35	8	22.9%	18	0	0.0%
Group 5	21	2	9.5%	15	1	6.7%
Total	135	26	19.3%	81	8	9.9%

G.3.3.3. Algebra & Algebraic Thinking (AA) vs. Statistics & Probability (SP)

Table A 54: Frequency of teachers in each percentile group for MQI component scores comparisons between AA and SP: Year One and Year Two

	MQI_Y1_Y2_AA_SP (AA as reference group)					
	Num in each group	AA1_SP1	% Unchanged	Num in each group	AA2_SP2	%Unchanged
Group 1	10	1	10.0%	10	4	40.0%
Group 2	11	1	9.1%	11	3	27.3%
Group 3	11	2	18.2%	11	3	27.3%
Group 4	11	0	0.0%	11	1	9.1%
Group 5	10	1	10.0%	10	3	30.0%
Total	53	5	9.4%	53	14	26.4%

Table A 55: Frequency of teachers in each percentile group for MQI simple average scores comparisons between AA and SP: Year One and Year Two

	MQI_AA_SP (AA as reference group)		
	Num in each group	Both Years	% Unchanged
Group 1	8	4	50.0%
Group 2	13	3	23.1%
Group 3	12	1	8.3%
Group 4	10	1	10.0%
Group 5	10	0	0.0%
Total	58	9	15.5%

APPENDIX H:

FREQUENCY TABLE FOR CHANGE IN PERCENTILE GROUPS

H.1. Generalist Teachers' Change in Ranks by Instrument (ELA as the baseline)

Table A 56: Year One generalist teachers' change in ranks on FFT: The first component

Difference in percentile group for the first component in ELA and Math

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	68	15.5	15.5	15.5
	1	111	25.2	25.2	40.7
	2	83	18.9	18.9	59.5
	3	58	13.2	13.2	72.7
	4	52	11.8	11.8	84.5
	5	23	5.2	5.2	89.8
	6	23	5.2	5.2	95.0
	7	19	4.3	4.3	99.3
	8	1	.2	.2	99.5
	9	2	.5	.5	100.0
	Total	440	100.0	100.0	

Table A 57: Year One generalist teachers' change in ranks on FFT: The second component

Difference in percentile group for the second component in ELA and Math

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	69	15.7	15.7	15.7
	1	109	24.8	24.8	40.5
	2	90	20.5	20.5	60.9
	3	66	15.0	15.0	75.9
	4	45	10.2	10.2	86.1
	5	28	6.4	6.4	92.5
	6	18	4.1	4.1	96.6
	7	11	2.5	2.5	99.1
	8	4	.9	.9	100.0
	Total	440	100.0	100.0	

Table A 58: Year One generalist teachers' change in ranks on FFT: The simple average

Difference in percentile group between ELA and math averages

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	77	17.5	17.5	17.5
	1	100	22.7	22.7	40.2
	2	92	20.9	20.9	61.1
	3	61	13.9	13.9	75.0
	4	43	9.8	9.8	84.8
	5	35	8.0	8.0	92.7
	6	13	3.0	3.0	95.7
	7	15	3.4	3.4	99.1
	8	3	.7	.7	99.8
	9	1	.2	.2	100.0
	Total	440	100.0	100.0	

Table A 59: Year Two generalist teachers' change in ranks on FFT: The first component

Difference in percentile group for the first component in ELA and Math

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	54	17.3	17.3	17.3
	1	76	24.3	24.3	41.5
	2	65	20.8	20.8	62.3
	3	38	12.1	12.1	74.4
	4	38	12.1	12.1	86.6
	5	19	6.1	6.1	92.7
	6	10	3.2	3.2	95.8
	7	9	2.9	2.9	98.7
	8	4	1.3	1.3	100.0
	Total	313	100.0	100.0	

Table A 60: Year Two generalist teachers' change in ranks on FFT: The second component

Difference in percentile group for the second component in ELA and Math

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	45	14.4	14.4	14.4
	1	90	28.8	28.8	43.1
	2	58	18.5	18.5	61.7
	3	49	15.7	15.7	77.3
	4	32	10.2	10.2	87.5
	5	22	7.0	7.0	94.6
	6	12	3.8	3.8	98.4
	7	3	1.0	1.0	99.4
	8	2	.6	.6	100.0
	Total	313	100.0	100.0	

Table A 61: Year Two generalist teachers' change in ranks on FFT: The simple average

Difference in percentile group between ELA and math averages

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	44	14.1	14.1	14.1
	1	90	28.8	28.8	42.8
	2	58	18.5	18.5	61.3
	3	46	14.7	14.7	76.0
	4	39	12.5	12.5	88.5
	5	15	4.8	4.8	93.3
	6	13	4.2	4.2	97.4
	7	6	1.9	1.9	99.4
	8	2	.6	.6	100.0
	Total	313	100.0	100.0	

Table A 62: Year One generalist teachers' change in ranks on CLASS: The first component

Difference in percentile group of the first component for ELA vs. Math

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	74	16.8	16.8	16.8
	1	109	24.8	24.8	41.6
	2	84	19.1	19.1	60.7
	3	72	16.4	16.4	77.0
	4	44	10.0	10.0	87.0
	5	30	6.8	6.8	93.9
	6	17	3.9	3.9	97.7
	7	8	1.8	1.8	99.5
	8	1	.2	.2	99.8
	9	1	.2	.2	100.0
	Total	440	100.0	100.0	

Table A 63: Year Two generalist teachers' change in ranks on CLASS: The second component

Difference in percentile group of the second component for ELA vs. Math

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	85	19.3	19.3	19.3
	1	121	27.5	27.5	46.8
	2	83	18.9	18.9	65.7
	3	76	17.3	17.3	83.0
	4	38	8.6	8.6	91.6
	5	18	4.1	4.1	95.7
	6	11	2.5	2.5	98.2
	7	7	1.6	1.6	99.8
	8	1	.2	.2	100.0
	Total	440	100.0	100.0	

Table A 64: Year Two generalist teachers' change in ranks on CLASS: The simple average

Difference in percentile group of simple average for ELA vs. Math

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	83	18.9	18.9	18.9
	1	111	25.2	25.2	44.1
	2	82	18.6	18.6	62.7
	3	75	17.0	17.0	79.8
	4	49	11.1	11.1	90.9
	5	20	4.5	4.5	95.5
	6	11	2.5	2.5	98.0
	7	6	1.4	1.4	99.3
	8	3	.7	.7	100.0
	Total	440	100.0	100.0	

Table A 65: Year Two generalist teachers' change in ranks on CLASS: The first component

Difference in percentile group of the first component for ELA vs. Math

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	63	20.1	20.1	20.1
	1	86	27.5	27.5	47.6
	2	58	18.5	18.5	66.1
	3	27	8.6	8.6	74.8
	4	32	10.2	10.2	85.0
	5	25	8.0	8.0	93.0
	6	7	2.2	2.2	95.2
	7	10	3.2	3.2	98.4
	8	3	1.0	1.0	99.4
	9	2	.6	.6	100.0
	Total	313	100.0	100.0	

Table A 66: Year Two generalist teachers' change in ranks on CLASS: The second component

Difference in percentile group of the second component for ELA vs. Math

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	50	16.0	16.0	16.0
	1	95	30.4	30.4	46.3
	2	54	17.3	17.3	63.6
	3	43	13.7	13.7	77.3
	4	27	8.6	8.6	85.9
	5	23	7.3	7.3	93.3
	6	14	4.5	4.5	97.8
	7	4	1.3	1.3	99.0
	8	2	.6	.6	99.7
	9	1	.3	.3	100.0
	Total	313	100.0	100.0	

Table A 67: Year Two generalist teachers' change in ranks on CLASS: The simple average

Difference in percentile group of simple average for ELA vs. Math

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	56	17.9	17.9	17.9
	1	88	28.1	28.1	46.0
	2	66	21.1	21.1	67.1
	3	36	11.5	11.5	78.6
	4	29	9.3	9.3	87.9
	5	19	6.1	6.1	93.9
	6	7	2.2	2.2	96.2
	7	6	1.9	1.9	98.1
	8	4	1.3	1.3	99.4
	9	2	.6	.6	100.0
	Total	313	100.0	100.0	

Table A 68: Year Two generalist teachers' change in ranks from PLATO vs. MQI: The first component

Difference in percentile group between PLATO PC1 and MQI PC1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	63	14.7	14.7	14.7
	1	86	20.0	20.0	34.7
	2	73	17.0	17.0	51.6
	3	65	15.1	15.1	66.7
	4	47	10.9	10.9	77.7
	5	35	8.1	8.1	85.8
	6	29	6.7	6.7	92.6
	7	14	3.3	3.3	95.8
	8	14	3.3	3.3	99.1
	9	4	.9	.9	100.0
	Total	430	100.0	100.0	

Table A 69: Year Two generalist teachers' change in ranks from PLATO vs. MQI: The second component

Difference in percentile group between PLATO PC2 and MQI PC1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	48	11.2	11.2	11.2
	1	89	20.7	20.7	31.9
	2	72	16.7	16.7	48.6
	3	59	13.7	13.7	62.3
	4	53	12.3	12.3	74.7
	5	48	11.2	11.2	85.8
	6	27	6.3	6.3	92.1
	7	15	3.5	3.5	95.6
	8	12	2.8	2.8	98.4
	9	7	1.6	1.6	100.0
	Total	430	100.0	100.0	

Table A 70: Year Two generalist teachers' change in ranks from PLATO vs. MQI: The simple average

Difference in percentile group between PLATO average and MQI average

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	57	13.3	13.3	13.3
	1	81	18.8	18.8	32.1
	2	66	15.3	15.3	47.4
	3	83	19.3	19.3	66.7
	4	40	9.3	9.3	76.0
	5	43	10.0	10.0	86.0
	6	26	6.0	6.0	92.1
	7	16	3.7	3.7	95.8
	8	14	3.3	3.3	99.1
	9	4	.9	.9	100.0
	Total	430	100.0	100.0	

Table A 71: Year Two generalist teachers' change in ranks from PLATO vs. MQI: The first component

Difference in percentile group between MQI1 and PLATO1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	46	14.8	14.8	14.8
	1	64	20.6	20.6	35.5
	2	55	17.7	17.7	53.2
	3	43	13.9	13.9	67.1
	4	37	11.9	11.9	79.0
	5	22	7.1	7.1	86.1
	6	17	5.5	5.5	91.6
	7	13	4.2	4.2	95.8
	8	10	3.2	3.2	99.0
	9	3	1.0	1.0	100.0
	Total	310	100.0	100.0	

Table A 72: Year Two generalist teachers' change in ranks from PLATO vs. MQI: The second component

Difference in percentile group between MQI1 and PLATO2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	30	9.7	9.7	9.7
	1	65	21.0	21.0	30.6
	2	51	16.5	16.5	47.1
	3	47	15.2	15.2	62.3
	4	35	11.3	11.3	73.5
	5	30	9.7	9.7	83.2
	6	26	8.4	8.4	91.6
	7	15	4.8	4.8	96.5
	8	9	2.9	2.9	99.4
	9	2	.6	.6	100.0
	Total	310	100.0	100.0	

Table A 73: Year Two generalist teachers' change in ranks from PLATO vs. MQI: The simple average

Difference in percentile group between ELA and Math

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	33	10.6	10.6	10.6
	1	55	17.7	17.7	28.4
	2	70	22.6	22.6	51.0
	3	39	12.6	12.6	63.5
	4	43	13.9	13.9	77.4
	5	28	9.0	9.0	86.5
	6	21	6.8	6.8	93.2
	7	9	2.9	2.9	96.1
	8	9	2.9	2.9	99.0
	9	3	1.0	1.0	100.0
	Total	310	100.0	100.0	

H.2. Mathematics Teachers' Changes in Ranks between Across Different Instruments

Table A 74: Year One mathematics teachers' change in ranks for FFT vs. CLASS: The first component

Difference in percentile group between FFT PC1 and CLASS PC1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	227	23.2	23.2	23.2
	1	272	27.8	27.8	51.0
	2	200	20.4	20.4	71.5
	3	145	14.8	14.8	86.3
	4	78	8.0	8.0	94.3
	5	33	3.4	3.4	97.6
	6	12	1.2	1.2	98.9
	7	9	.9	.9	99.8
	8	2	.2	.2	100.0
	Total	978	100.0	100.0	

Table A 75: Year One mathematics teachers' change in ranks for FFT vs. CLASS: The second component

Difference in percentile group between FFT PC2 and CLASS PC2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	240	24.5	24.5	24.5
	1	326	33.3	33.3	57.9
	2	191	19.5	19.5	77.4
	3	107	10.9	10.9	88.3
	4	64	6.5	6.5	94.9
	5	38	3.9	3.9	98.8
	6	10	1.0	1.0	99.8
	8	1	.1	.1	99.9
	9	1	.1	.1	100.0
	Total	978	100.0	100.0	

Table A 76: Year One mathematics teachers' change in ranks for FFT vs. CLASS: The simple average

Difference in percentile group between FFT average and CLASS average on mathematics lessons

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	235	24.0	24.0	24.0
	1	287	29.3	29.3	53.4
	2	199	20.3	20.3	73.7
	3	137	14.0	14.0	87.7
	4	63	6.4	6.4	94.2
	5	39	4.0	4.0	98.2
	6	7	.7	.7	98.9
	7	9	.9	.9	99.8
	8	2	.2	.2	100.0
	Total	978	100.0	100.0	

Table A 77: Year Two mathematics teachers' change in ranks for FFT vs. CLASS: The first component

Difference in percentile group between FFT PC1 and CLASS PC1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	188	24.4	24.4	24.4
	1	210	27.2	27.2	51.6
	2	154	19.9	19.9	71.5
	3	104	13.5	13.5	85.0
	4	58	7.5	7.5	92.5
	5	27	3.5	3.5	96.0
	6	22	2.8	2.8	98.8
	7	8	1.0	1.0	99.9
	9	1	.1	.1	100.0
	Total	772	100.0	100.0	

Table A 78: Year Two mathematics teachers' change in ranks for FFT vs. CLASS: The second component

Difference in percentile group between FFT PC2 and CLASS PC2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	177	22.9	22.9	22.9
	1	248	32.1	32.1	55.1
	2	159	20.6	20.6	75.6
	3	92	11.9	11.9	87.6
	4	52	6.7	6.7	94.3
	5	24	3.1	3.1	97.4
	6	13	1.7	1.7	99.1
	7	6	.8	.8	99.9
	8	1	.1	.1	100.0
	Total	772	100.0	100.0	

Table A 79: Year Two mathematics teachers' change in ranks for FFT vs. CLASS: The simple average

Difference in percentile group between FFT average and CLASS average

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	173	22.4	22.4	22.4
	1	234	30.3	30.3	52.7
	2	158	20.5	20.5	73.2
	3	102	13.2	13.2	86.4
	4	54	7.0	7.0	93.4
	5	27	3.5	3.5	96.9
	6	19	2.5	2.5	99.4
	7	4	.5	.5	99.9
	8	1	.1	.1	100.0
	Total	772	100.0	100.0	

Table A 80: Year One mathematics teachers' change in ranks for FFT vs. MQI: The first component

Difference in percentile group between FFT PC1 and MQI PC1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	135	13.9	13.9	13.9
	1	234	24.1	24.1	38.0
	2	178	18.3	18.3	56.3
	3	145	14.9	14.9	71.3
	4	102	10.5	10.5	81.8
	5	87	9.0	9.0	90.7
	6	49	5.0	5.0	95.8
	7	24	2.5	2.5	98.2
	8	15	1.5	1.5	99.8
	9	2	.2	.2	100.0
	Total	971	100.0	100.0	

Table A 81: Year One mathematics teachers' change in ranks for FFT vs. MQI: The simple average

Difference in percentile group between FFT average and MQI average

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	138	14.2	14.2	14.2
	1	230	23.7	23.7	37.9
	2	167	17.2	17.2	55.1
	3	139	14.3	14.3	69.4
	4	98	10.1	10.1	79.5
	5	91	9.4	9.4	88.9
	6	56	5.8	5.8	94.6
	7	34	3.5	3.5	98.1
	8	15	1.5	1.5	99.7
	9	3	.3	.3	100.0
	Total	971	100.0	100.0	

Table A 82: Year Two mathematics teachers' change in ranks from FFT vs. MQI: The first component

Difference in percentile group between FFT PC1 and MQI PC1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	103	13.4	13.4	13.4
	1	185	24.0	24.0	37.4
	2	134	17.4	17.4	54.8
	3	110	14.3	14.3	69.1
	4	82	10.6	10.6	79.7
	5	70	9.1	9.1	88.8
	6	40	5.2	5.2	94.0
	7	21	2.7	2.7	96.8
	8	18	2.3	2.3	99.1
	9	7	.9	.9	100.0
	Total	770	100.0	100.0	

Table A 83: Year Two mathematics teachers' change in ranks for FFT vs. MQI: The simple average

Difference in percentile group between FFT average and MQI average

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	95	12.3	12.3	12.3
	1	182	23.6	23.6	36.0
	2	136	17.7	17.7	53.6
	3	113	14.7	14.7	68.3
	4	82	10.6	10.6	79.0
	5	64	8.3	8.3	87.3
	6	37	4.8	4.8	92.1
	7	37	4.8	4.8	96.9
	8	16	2.1	2.1	99.0
	9	8	1.0	1.0	100.0
	Total	770	100.0	100.0	

Table A 84: Year One mathematics teachers' change in ranks for CLASS vs. MQI: The first component

Difference in percentile group between CLASS PC1 and MQI PC1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	160	16.5	16.5	16.5
	1	250	25.7	25.7	42.2
	2	179	18.4	18.4	60.7
	3	136	14.0	14.0	74.7
	4	93	9.6	9.6	84.2
	5	69	7.1	7.1	91.3
	6	46	4.7	4.7	96.1
	7	25	2.6	2.6	98.7
	8	9	.9	.9	99.6
	9	4	.4	.4	100.0
	Total	971	100.0	100.0	

Table A 85: Year One mathematics teachers' change in ranks for CLASS vs. MQI: The simple average

Difference in percentile group between CLASS average and MQI average

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	134	13.8	13.8	13.8
	1	233	24.0	24.0	37.8
	2	181	18.6	18.6	56.4
	3	136	14.0	14.0	70.4
	4	114	11.7	11.7	82.2
	5	78	8.0	8.0	90.2
	6	47	4.8	4.8	95.1
	7	24	2.5	2.5	97.5
	8	21	2.2	2.2	99.7
	9	3	.3	.3	100.0
	Total	971	100.0	100.0	

Table A 86: Year Two mathematics teachers' change in ranks for CLASS vs. MQI: The first component

Difference in percentile group between CLASS PC1 and MQI PC1 for mathematics teachers

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	141	18.3	18.3	18.3
	1	162	21.0	21.0	39.4
	2	159	20.6	20.6	60.0
	3	98	12.7	12.7	72.7
	4	84	10.9	10.9	83.6
	5	58	7.5	7.5	91.2
	6	35	4.5	4.5	95.7
	7	23	3.0	3.0	98.7
	8	7	.9	.9	99.6
	9	3	.4	.4	100.0
	Total	770	100.0	100.0	

Table A 87: Year Two mathematics teachers' change in ranks for CLASS vs. MQI: The simple average

Difference in percentile group of simple average between CLASS and MQI for mathematics teachers

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	96	12.5	12.5	12.5
	1	200	26.0	26.0	38.4
	2	133	17.3	17.3	55.7
	3	104	13.5	13.5	69.2
	4	98	12.7	12.7	81.9
	5	56	7.3	7.3	89.2
	6	45	5.8	5.8	95.1
	7	27	3.5	3.5	98.6
	8	8	1.0	1.0	99.6
	9	3	.4	.4	100.0
	Total	770	100.0	100.0	

H.3. Mathematics Teachers' Changes in Ranks between Pairs of Subject Areas within Mathematics in Different Instrument

FFT Year One: Algebra & Algebraic Thinking (AA) vs. Numbers & Operations (NO)

Table A 88: Year One mathematics teachers' change in ranks between AA & NO on FFT: The first component

Difference in percentile group for the first component between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	34	14.8	14.8	14.8
	1	52	22.6	22.6	37.4
	2	38	16.5	16.5	53.9
	3	27	11.7	11.7	65.7
	4	32	13.9	13.9	79.6
	5	22	9.6	9.6	89.1
	6	13	5.7	5.7	94.8
	7	6	2.6	2.6	97.4
	8	4	1.7	1.7	99.1
	9	2	.9	.9	100.0
	Total	230	100.0	100.0	

Table A 89: Year One mathematics teachers' change in ranks between AA & NO on FFT: The second component

Difference in percentile group for the first component between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	33	14.3	14.3	14.3
	1	59	25.7	25.7	40.0
	2	39	17.0	17.0	57.0
	3	30	13.0	13.0	70.0
	4	29	12.6	12.6	82.6
	5	15	6.5	6.5	89.1
	6	10	4.3	4.3	93.5
	7	10	4.3	4.3	97.8
	8	4	1.7	1.7	99.6
	9	1	.4	.4	100.0
	Total	230	100.0	100.0	

Table A 90: Year One mathematics teachers' change in ranks between AA & NO on FFT: The simple average

Difference in percentile group for simple average between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	38	16.5	16.5	16.5
	1	50	21.7	21.7	38.3
	2	40	17.4	17.4	55.7
	3	26	11.3	11.3	67.0
	4	26	11.3	11.3	78.3
	5	24	10.4	10.4	88.7
	6	12	5.2	5.2	93.9
	7	9	3.9	3.9	97.8
	8	3	1.3	1.3	99.1
	9	2	.9	.9	100.0
	Total	230	100.0	100.0	

FFT Year One: Geometry (G) vs. Numbers & Operations (NO)

Table A 91: Year One mathematics teachers' change in ranks between G & NO on FFT: The first component

Difference in percentile group for the first component between Geometry and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	13	9.6	9.6	9.6
	1	35	25.9	25.9	35.6
	2	22	16.3	16.3	51.9
	3	22	16.3	16.3	68.1
	4	15	11.1	11.1	79.3
	5	9	6.7	6.7	85.9
	6	8	5.9	5.9	91.9
	7	8	5.9	5.9	97.8
	8	3	2.2	2.2	100.0
	Total	135	100.0	100.0	

Table A 92: Year One mathematics teachers' change in ranks between G & NO on FFT: The second component

Difference in percentile group for the second component between Geometry and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	23	17.0	17.0	17.0
	1	27	20.0	20.0	37.0
	2	27	20.0	20.0	57.0
	3	14	10.4	10.4	67.4
	4	17	12.6	12.6	80.0
	5	9	6.7	6.7	86.7
	6	7	5.2	5.2	91.9
	7	7	5.2	5.2	97.0
	8	3	2.2	2.2	99.3
	9	1	.7	.7	100.0
	Total	135	100.0	100.0	

Table A 93: Year One mathematics teachers' change in ranks between G & NO on FFT: The simple average

Difference in percentile group for average between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	37	27.4	27.4	27.4
	1	49	36.3	36.3	63.7
	2	27	20.0	20.0	83.7
	3	16	11.9	11.9	95.6
	4	6	4.4	4.4	100.0
	Total	135	100.0	100.0	

FFT Year Two: Algebra & Algebraic Thinking (AA) vs. Numbers & Operations (NO)

Table A 94: Year Two mathematics teachers' change in ranks between AA & NO on FFT: The first component

Difference in percentile group for the first component between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	36	20.6	20.6	20.6
	1	44	25.1	25.1	45.7
	2	25	14.3	14.3	60.0
	3	22	12.6	12.6	72.6
	4	16	9.1	9.1	81.7
	5	11	6.3	6.3	88.0
	6	9	5.1	5.1	93.1
	7	11	6.3	6.3	99.4
	8	1	.6	.6	100.0
	Total	175	100.0	100.0	

Table A 95: Year Two mathematics teachers' change in ranks between AA & NO on FFT: The second component

Difference in percentile group for the second component between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	33	18.9	18.9	18.9
	1	47	26.9	26.9	45.7
	2	26	14.9	14.9	60.6
	3	22	12.6	12.6	73.1
	4	17	9.7	9.7	82.9
	5	14	8.0	8.0	90.9
	6	6	3.4	3.4	94.3
	7	9	5.1	5.1	99.4
	8	1	.6	.6	100.0
	Total	175	100.0	100.0	

Table A 96: Year Two mathematics teachers' change in ranks between AA & NO on FFT: The simple average

Difference in percentile group for simple average between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	37	21.1	21.1	21.1
	1	42	24.0	24.0	45.1
	2	29	16.6	16.6	61.7
	3	20	11.4	11.4	73.1
	4	14	8.0	8.0	81.1
	5	12	6.9	6.9	88.0
	6	10	5.7	5.7	93.7
	7	10	5.7	5.7	99.4
	8	1	.6	.6	100.0
	Total	175	100.0	100.0	

FFT Year Two: Numbers & Operations (NO) vs. Geometry (G)

Table A 97: Year Two mathematics teachers' change in ranks between AA & NO on FFT: The simple average

Difference in percentile group for the first component between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	27	32.1	32.1	32.1
	1	20	23.8	23.8	56.0
	2	25	29.8	29.8	85.7
	3	10	11.9	11.9	97.6
	4	2	2.4	2.4	100.0
	Total	84	100.0	100.0	

Table A 98: Year Two mathematics teachers' change in ranks between G & NO on FFT: The second component

Difference in percentile group for the second component between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	21	25.0	25.0	25.0
	1	30	35.7	35.7	60.7
	2	22	26.2	26.2	86.9
	3	9	10.7	10.7	97.6
	4	2	2.4	2.4	100.0
	Total	84	100.0	100.0	

Table A 99: Year Two mathematics teachers' change in ranks between G & NO on FFT: The first component

Difference in percentile group for simple average between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	25	29.8	29.8	29.8
	1	23	27.4	27.4	57.1
	2	25	29.8	29.8	86.9
	3	10	11.9	11.9	98.8
	4	1	1.2	1.2	100.0
	Total	84	100.0	100.0	

FFT Year One & Two: Algebra & Algebraic Thinking (AA) vs. Statistics & Probability (SP)

Table A 100: Year One & Two mathematics teachers' change in ranks between AA & SP on FFT: The first component

Difference in percentile group for the first components between Algebra & Algebraic Thinking and Statistics & Probability

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	21	36.2	36.2	36.2
	1	16	27.6	27.6	63.8
	2	11	19.0	19.0	82.8
	3	7	12.1	12.1	94.8
	4	3	5.2	5.2	100.0
	Total	58	100.0	100.0	

Table A 101: Year One & Two mathematics teachers' change in ranks between AA & SP on FFT: The second component

Difference in percentile group for the second components between Algebra & Algebraic Thinking and Statistics & Probability

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	19	32.8	32.8	32.8
	1	16	27.6	27.6	60.3
	2	16	27.6	27.6	87.9
	3	4	6.9	6.9	94.8
	4	3	5.2	5.2	100.0
	Total	58	100.0	100.0	

Table A 102: Year One & Two mathematics teachers' change in ranks between AA & SP on FFT: The simple average

Difference in percentile group between Algebra & Algebraic Thinking and Statistics & Probability

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	17	30.4	30.4	30.4
	1	17	30.4	30.4	60.7
	2	12	21.4	21.4	82.1
	3	9	16.1	16.1	98.2
	4	1	1.8	1.8	100.0
	Total	56	100.0	100.0	

CLASS Year One: Algebra & Algebraic Thinking (AA) vs. Numbers & Operations (NO)

Table A 103: Year One mathematics teachers' change in ranks between AA & NO on CLASS: The first component

Difference in percentile group of the first component between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	42	18.2	18.2	18.2
	1	43	18.6	18.6	36.8
	2	48	20.8	20.8	57.6
	3	27	11.7	11.7	69.3
	4	32	13.9	13.9	83.1
	5	16	6.9	6.9	90.0
	6	15	6.5	6.5	96.5
	7	2	.9	.9	97.4
	8	4	1.7	1.7	99.1
	9	2	.9	.9	100.0
	Total	231	100.0	100.0	

Table A 104: Year One mathematics teachers' change in ranks between AA & NO on CLASS: The second component

Difference in percentile group of the second component between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	39	16.9	16.9	16.9
	1	45	19.5	19.5	36.4
	2	41	17.7	17.7	54.1
	3	35	15.2	15.2	69.3
	4	25	10.8	10.8	80.1
	5	24	10.4	10.4	90.5
	6	14	6.1	6.1	96.5
	7	5	2.2	2.2	98.7
	8	2	.9	.9	99.6
	9	1	.4	.4	100.0
	Total	231	100.0	100.0	

Table A 105: Year One mathematics teachers' change in ranks between AA & NO on CLASS: The simple average

Difference in percentile group of simple average between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	41	17.7	17.7	17.7
	1	52	22.5	22.5	40.3
	2	36	15.6	15.6	55.8
	3	31	13.4	13.4	69.3
	4	31	13.4	13.4	82.7
	5	17	7.4	7.4	90.0
	6	10	4.3	4.3	94.4
	7	9	3.9	3.9	98.3
	8	2	.9	.9	99.1
	9	2	.9	.9	100.0
	Total	231	100.0	100.0	

CLASS Year One: Geometry (G) vs. Numbers & Operations (NO)

Table A 106: Year One mathematics teachers' change in ranks between G & NO on CLASS: The first component

Difference in percentile group of the first component between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	17	12.7	12.7	12.7
	1	28	20.9	20.9	33.6
	2	29	21.6	21.6	55.2
	3	17	12.7	12.7	67.9
	4	13	9.7	9.7	77.6
	5	14	10.4	10.4	88.1
	6	8	6.0	6.0	94.0
	7	5	3.7	3.7	97.8
	8	3	2.2	2.2	100.0
	Total	134	100.0	100.0	

Table A 107: Year One mathematics teachers' change in ranks between G & NO on CLASS: The second component

Difference in percentile group of the second component between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	17	12.7	12.7	12.7
	1	36	26.9	26.9	39.6
	2	25	18.7	18.7	58.2
	3	20	14.9	14.9	73.1
	4	12	9.0	9.0	82.1
	5	11	8.2	8.2	90.3
	6	5	3.7	3.7	94.0
	7	5	3.7	3.7	97.8
	8	3	2.2	2.2	100.0
	Total	134	100.0	100.0	

Table A 108: Year One mathematics teachers' change in ranks between G & NO on CLASS: The simple average

Difference in percentile group of the simple average between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	18	13.4	13.4	13.4
	1	27	20.1	20.1	33.6
	2	28	20.9	20.9	54.5
	3	19	14.2	14.2	68.7
	4	15	11.2	11.2	79.9
	5	9	6.7	6.7	86.3
	6	13	9.7	9.7	96.3
	7	3	2.2	2.2	98.5
	8	2	1.5	1.5	100.0
	Total	134	100.0	100.0	

CLASS Year Two: Algebra & Algebraic Thinking vs. Numbers & Operations

Table A 109: Year Two mathematics teachers' change in ranks between AA & NO on CLASS: The first component

Difference in percentile group of the first component between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	35	20.0	20.0	20.0
	1	41	23.4	23.4	43.4
	2	28	16.0	16.0	59.4
	3	27	15.4	15.4	74.9
	4	17	9.7	9.7	84.6
	5	16	9.1	9.1	93.7
	6	8	4.6	4.6	98.3
	8	3	1.7	1.7	100.0
	Total	175	100.0	100.0	

Table A 110: Year Two mathematics teachers' change in ranks between AA & NO on CLASS: The second component

Difference in percentile group of the second component between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	30	17.1	17.1	17.1
	1	48	27.4	27.4	44.6
	2	40	22.9	22.9	67.4
	3	15	8.6	8.6	76.0
	4	17	9.7	9.7	85.7
	5	14	8.0	8.0	93.7
	6	5	2.9	2.9	96.6
	7	4	2.3	2.3	98.9
	8	2	1.1	1.1	100.0
	Total	175	100.0	100.0	

Table A 111: Year Two mathematics teachers' change in ranks between AA & NO on CLASS: The simple average

Difference in percentile group of simple average between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	32	18.3	18.3	18.3
	1	43	24.6	24.6	42.9
	2	31	17.7	17.7	60.6
	3	28	16.0	16.0	76.6
	4	17	9.7	9.7	86.3
	5	13	7.4	7.4	93.7
	6	6	3.4	3.4	97.1
	7	3	1.7	1.7	98.9
	8	2	1.1	1.1	100.0
	Total	175	100.0	100.0	

CLASS Year Two: Geometry (G) vs. Numbers & Operations (NO)

Table A 112: Year Two mathematics teachers' change in ranks between G & NO on CLASS: The first component

Difference in percentile group of the first component between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	25	29.8	29.8	29.8
	1	31	36.9	36.9	66.7
	2	15	17.9	17.9	84.5
	3	11	13.1	13.1	97.6
	4	2	2.4	2.4	100.0
	Total	84	100.0	100.0	

Table A 113: Year Two mathematics teachers' change in ranks between G & NO on CLASS: The second component

Difference in percentile group of the second component between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	21	25.0	25.0	25.0
	1	32	38.1	38.1	63.1
	2	17	20.2	20.2	83.3
	3	10	11.9	11.9	95.2
	4	4	4.8	4.8	100.0
	Total	84	100.0	100.0	

Table A 114: Year Two mathematics teachers' change in ranks between G & NO on CLASS: The simple average

Difference in percentile group of the simple average between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	17	20.2	20.2	20.2
	1	36	42.9	42.9	63.1
	2	17	20.2	20.2	83.3
	3	13	15.5	15.5	98.8
	4	1	1.2	1.2	100.0
	Total	84	100.0	100.0	

CLASS Year One and Year Two: Algebra & Algebraic Thinking vs. Statistics & Probability

Table A 115: Year One and Year Two mathematics teachers' change in ranks between AA & SP on CLASS: The first component

Difference in percentile group of the first component between Algebra & Algebraic Thinking and Statistics & Probability

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	16	28.6	28.6	28.6
	1	23	41.1	41.1	69.6
	2	9	16.1	16.1	85.7
	3	7	12.5	12.5	98.2
	4	1	1.8	1.8	100.0
	Total	56	100.0	100.0	

Table A 116: Year One and Year Two mathematics teachers' change in ranks between AA & SP on CLASS: The second component

Difference in percentile group of the second component between Algebra & Algebraic Thinking and Statistics & Probability

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	7	12.5	12.5	12.5
	1	17	30.4	30.4	42.9
	2	21	37.5	37.5	80.4
	3	7	12.5	12.5	92.9
	4	4	7.1	7.1	100.0
	Total	56	100.0	100.0	

Table A 117: Year One and Year Two mathematics teachers' change in ranks between AA & SP on CLASS: The simple average

Difference in percentile group of the simple average t between Algebra & Algebraic Thinking and Statistics & Probability

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	17	30.4	30.4	30.4
	1	18	32.1	32.1	62.5
	2	11	19.6	19.6	82.1
	3	9	16.1	16.1	98.2
	4	1	1.8	1.8	100.0
	Total	56	100.0	100.0	

MQI Year One: Numbers & Operations (NO) vs. Algebra & Algebraic Thinking (AA)

Table A 118: Year One mathematics teachers' change in ranks between AA & NO on MQI: The first component

Difference in percentile group of the first component between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	34	15.4	15.4	15.4
	1	46	20.8	20.8	36.2
	2	35	15.8	15.8	52.0
	3	34	15.4	15.4	67.4
	4	21	9.5	9.5	76.9
	5	18	8.1	8.1	85.1
	6	11	5.0	5.0	90.0
	7	13	5.9	5.9	95.9
	8	6	2.7	2.7	98.6
	9	3	1.4	1.4	100.0
	Total	221	100.0	100.0	

Table A 119: Year One mathematics teachers' change in ranks between AA & NO on MQI: The second component

Difference in percentile group of the second component between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	34	15.4	15.4	15.4
	1	37	16.7	16.7	32.1
	2	42	19.0	19.0	51.1
	3	25	11.3	11.3	62.4
	4	19	8.6	8.6	71.0
	5	29	13.1	13.1	84.2
	6	17	7.7	7.7	91.9
	7	9	4.1	4.1	95.9
	8	6	2.7	2.7	98.6
	9	3	1.4	1.4	100.0
	Total	221	100.0	100.0	

Table A 120: Year One mathematics teachers' change in ranks between AA & NO on MQI: The simple average

Difference in percentile group of the simple average between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	24	10.9	10.9	10.9
	1	48	21.7	21.7	32.6
	2	32	14.5	14.5	47.1
	3	36	16.3	16.3	63.3
	4	35	15.8	15.8	79.2
	5	15	6.8	6.8	86.0
	6	10	4.5	4.5	90.5
	7	11	5.0	5.0	95.5
	8	6	2.7	2.7	98.2
	9	4	1.8	1.8	100.0
	Total	221	100.0	100.0	

MQI Year One: Numbers & Operations (NO) vs. Geometry (G)

Table A 121: Year One mathematics teachers' change in ranks between G & NO on MQI: The first component

Difference in percentile group for the first component between Geometry and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	12	9.6	9.6	9.6
	1	24	19.2	19.2	28.8
	2	25	20.0	20.0	48.8
	3	17	13.6	13.6	62.4
	4	17	13.6	13.6	76.0
	5	13	10.4	10.4	86.4
	6	4	3.2	3.2	89.6
	7	6	4.8	4.8	94.4
	8	6	4.8	4.8	99.2
	9	1	.8	.8	100.0
	Total	125	100.0	100.0	

Table A 122: Year One mathematics teachers' change in ranks between G & NO on MQI: The second component

Difference in percentile group for the second component between Geometry and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	18	14.4	14.4	14.4
	1	25	20.0	20.0	34.4
	2	22	17.6	17.6	52.0
	3	18	14.4	14.4	66.4
	4	14	11.2	11.2	77.6
	5	14	11.2	11.2	88.8
	6	5	4.0	4.0	92.8
	7	6	4.8	4.8	97.6
	8	32	1.6	1.6	99.2
	9	1	.8	.8	100.0
	Total	125	100.0	100.0	

Table A 123: Year One mathematics teachers' change in ranks between G & NO on MQI: The simple average

Difference in percentile group for average between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	26	20.8	20.8	20.8
	1	50	40.0	40.0	60.8
	2	31	24.8	24.8	85.6
	3	12	9.6	9.6	95.2
	4	6	4.8	4.8	100.0
	Total	125	100.0	100.0	

MQI Year Two: Algebra & Algebraic Thinking (AA) vs. Numbers & Operations (NO)

Table A 124: Year One mathematics teachers' change in ranks between AA & NO on MQI: The first component

Difference in percentile group for the first component between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	21	12.3	12.3	12.3
	1	42	24.6	24.6	36.8
	2	27	15.8	15.8	52.6
	3	18	10.5	10.5	63.2
	4	22	12.9	12.9	76.0
	5	14	8.2	8.2	84.2
	6	12	7.0	7.0	91.2
	7	5	2.9	2.9	94.2
	8	6	3.5	3.5	97.7
	9	4	2.3	2.3	100.0
	Total	171	100.0	100.0	

Table A 125: Year One mathematics teachers' change in ranks between AA & NO on MQI: The second component

Difference in percentile group for the second component between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	17	9.9	9.9	9.9
	1	32	18.7	18.7	28.7
	2	34	19.9	19.9	48.5
	3	25	14.6	14.6	63.2
	4	12	7.0	7.0	70.2
	5	17	9.9	9.9	80.1
	6	15	8.8	8.8	88.9
	7	10	5.8	5.8	94.7
	8	6	3.5	3.5	98.2
	9	3	1.8	1.8	100.0
	Total	171	100.0	100.0	

Table A 126: Year One mathematics teachers' change in ranks between AA & NO on MQI: The simple average

Difference in percentile group of simple average between Algebra & Algebraic Thinking and Numbers & Operations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	26	15.2	15.2	15.2
	1	30	17.5	17.5	32.7
	2	30	17.5	17.5	50.3
	3	20	11.7	11.7	62.0
	4	22	12.9	12.9	74.9
	5	15	8.8	8.8	83.6
	6	6	3.5	3.5	87.1
	7	15	8.8	8.8	95.9
	8	6	3.5	3.5	99.4
	9	1	.6	.6	100.0
	Total	171	100.0	100.0	

MQI Year Two: Numbers & Operations vs. Geometry

Table A 127: Year Two mathematics teachers' change in ranks between G & NO on MQI: The first component

Difference in percentile group for the first component between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	15	18.5	18.5	18.5
	1	21	25.9	25.9	44.4
	2	21	25.9	25.9	70.4
	3	16	19.8	19.8	90.1
	4	8	9.9	9.9	100.0
	Total	81	100.0	100.0	

Table A 128: Year Two mathematics teachers' change in ranks between G & NO on MQI: The second component

Difference in percentile group for the second component between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	14	17.3	17.3	17.3
	1	30	37.0	37.0	54.3
	2	12	14.8	14.8	69.1
	3	15	18.5	18.5	87.7
	4	10	12.3	12.3	100.0
	Total	81	100.0	100.0	

Table A 129: Year Two mathematics teachers' change in ranks between G & NO on MQI: The simple average

Difference in percentile group of simple average between Numbers & Operations and Geometry

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	8	9.9	9.9	9.9
	1	27	33.3	33.3	43.2
	2	22	27.2	27.2	70.4
	3	15	18.5	18.5	88.9
	4	9	11.1	11.1	100.0
	Total	81	100.0	100.0	

MQI Year One and Year Two: Algebra & Algebraic Thinking vs. Statistics & Probability

Table A 130: Year One and Year Two mathematics teachers' change in ranks between AA & SP on MQI: The first component

Difference in percentile group of the first component between Algebra & Algebraic Thinking and Statistics & Probability

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	5	9.4	9.4	9.4
	1	21	39.6	39.6	49.1
	2	11	20.8	20.8	69.8
	3	11	20.8	20.8	90.6
	4	5	9.4	9.4	100.0
	Total	53	100.0	100.0	

Table A 131: Year One and Year Two mathematics teachers' change in ranks between AA & SP on MQI: The second component

Difference in percentile group of the second component between Algebra & Algebraic Thinking and Statistics & Probability

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	14	26.4	26.4	26.4
	1	21	39.6	39.6	66.0
	2	12	22.6	22.6	88.7
	3	5	9.4	9.4	98.1
	4	1	1.9	1.9	100.0
	Total	53	100.0	100.0	

Table A 132: Year One and Year Two mathematics teachers' change in ranks between AA & SP on MQI: The simple average

Difference in percentile group of the simple average between Algebra & Algebraic Thinking and Statistics & Probability

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	9	17.0	17.0	17.0
	1	20	37.7	37.7	54.7
	2	14	26.4	26.4	81.1
	3	9	17.0	17.0	98.1
	4	1	1.9	1.9	100.0
	Total	53	100.0	100.0	

APPENDIX I:

ANOVA RESULTS FOR GENERALIST TEACHERS

Further investigations of other factors were performed on the significant cases among the list of comparisons regarding generalist teachers.

I.1. ANOVA Tables for Generalist Teachers on FFT

Table A 133: Summaries of P-values and effect sizes in ANOVA models for generalist teachers on FFT

	Year One			Year Two		
	FFT			FFT		
	PC1	PC2	SA	PC1	PC2	SA
Subject	0.060	0.849	0.252	0.660	0.828	0.804
Grade	0.506	0.543	0.547	0.363	0.097	0.259
Sub*Gr	0.218	0.219	0.199	0.438	0.302	0.389
Subject	0.008** (0.016)	0.808	0.128	0.001*** (0.039)	0.072	0.006** (0.024)
District	0.025** (0.025)	0.021* (0.026)	0.034* (0.024)	0.000*** (0.086)	0.000*** (0.064)	0.000*** (0.076)
Sub*Dist	0.225	0.139	0.160	0.767	0.565	0.701

Note: PC stands for principle component; SA stands for simple average. * means the effect is significant at the 0.05 level.

Table A 134: ANOVA with repeated measure on FFT for Year One generalist teachers: The influence of grade level and district

	GRADE					
Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
ANOVA 1: FFT Instruction scores as dependent variable						
A: SUBJECT	2.812	1	2.812	3.545	0.060	0.008
B: GRADE	2.745	2	1.373	0.683	0.506	0.003
A × B	2.428	1	1.214	1.531	0.218	0.007
Error(Within)	346.615	437	0.793			
Error(Between)	878.372	437	2.010			

Table A 134 (cont'd)

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
ANOVA 2: FFT Management scores as dependent variable						
A: SUBJECT	0.017	1	0.017	0.036	0.849	0.000
B: GRADE	1.815	2	0.908	0.612	0.543	0.003
A × B	1.418	2	0.709	1.549	0.219	0.007
Error(Within)	172.607	437	0.277			
Error(Between)	648.346	437	1.484			

Table A 134 (cont'd)

ANOVA 3: FFT Simple Average as dependent variable						
A: SUBJECT	0.042	1	0.042	1.314	0.252	0.003
B: GRADE	0.113	2	0.056	0.605	0.547	0.003
A × B	0.104	2	0.052	1.619	0.199	0.007
Error(Within)	14.084	437	0.032			
Error(Between)	40.726	437	0.093			

DISTRICT

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
ANOVA 4: FFT Instruction scores as dependent variable						
A: SUBJECT	5.675	1	5.675	7.165	0.008**	0.016
B: DISTRICT	22.187	4	5.547	2.809	0.025*	0.025
A × B	4.508	4	1.127	1.423	0.225	0.013
Error(Within)	344.535	435	0.792			
Error(Between)	858.930	435	1.975			
ANOVA 5: FFT Management Scores as dependent variable						
A: SUBJECT	0.027	1	0.027	0.059	0.808	0.000
B: DISTRICT	16.994	4	4.248	2.919	0.021*	0.026
A × B	3.180	4	0.795	1.745	0.139	0.016
Error(Within)	198.235	437	0.456			
Error(Between)	633.168	437	1.456			
ANOVA 6: FFT Simple Average as dependent variable						
A: SUBJECT	0.075	1	0.075	2.323	0.128	0.005
B: DISTRICT	0.965	4	0.241	2.632	0.034*	0.024
A × B	0.212	4	0.053	1.652	0.160	0.015
Error(Within)	13.976	435	0.032			
Error(Between)	39.874	435	0.092			

Note: * means significant at the 0.05 level. ** means significant at the 0.01 level.

Table A 135: ANOVA with repeated measure on FFT for Year Two Generalist Teachers

GRADE

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
ANOVA 1: FFT Instruction scores as dependent variable						
A: SUBJECT	0.172	1	0.172	0.194	0.660	0.001
B: GRADE	4.751	2	2.376	1.017	0.363	0.007
A × B	1.468	2	0.734	0.828	0.438	0.005
Error(Within)	274.668	310	0.886			
Error(Between)	724.884	310	2.335			
ANOVA 2: FFT Management scores as dependent variable						
A: SUBJECT	0.023	1	0.023	0.047	0.828	0.000
B: GRADE	7.775	2	3.887	2.352	0.097	0.015
A × B	1.195	2	0.598	1.201	0.302	0.008
Error(Within)	154.199	310	0.497			
Error(Between)	512.462	310	1.653			
ANOVA 3: FFT Simple Average as dependent variable						
A: SUBJECT	0.002	1	0.002	0.062	0.804	0.000
B: GRADE	0.288	2	0.144	1.355	0.259	0.009
A × B	0.067	2	0.034	0.946	0.389	0.006
Error(Within)	11.047	310	0.036			
Error(Between)	32.974	310	0.106			

DISTRICT

ANOVA 4: FFT Instruction as dependent variable						
A: SUBJECT	11.030	1	11.030	12.376	0.001***	0.039
B: DISTRICT	62.590	4	15.647	7.236	0.000***	0.086
A × B	1.632	4	0.408	0.458	0.767	0.006
Error(Within)	274.504	308	0.891			
Error(Between)	666.046	308	2.162			
ANOVA 5: FFT Management as dependent variable						
A: SUBJECT	1.629	1	1.629	3.260	0.072	0.010
B: DISTRICT	33.397	4	8.349	5.282	0.000***	0.064
A × B	1.479	4	0.370	0.740	0.565	0.010
Error(Within)	153.915	308	0.500			
Error(Between)	486.840	308	1.581			

Table A 135 (cont'd)

ANOVA 6: FFT Simple Average as dependent variable

A: SUBJECT	0.273	1	0.273	7.609	0.006**	0.024
B: DISTRICT	2.538	4	0.635	6.361	0.000***	0.076
A × B	0.078	4	0.020	0.547	0.701	0.007
Error(Within)	11.036	308	0.036			
Error(Between)	30.724	308	0.100			

Note: ** means significant at the 0.01 level, and *** means significant at the 0.001 level.

I.2. ANOVA Tables for General Teachers on CLASS

Table A 136: Summaries of P-values and effect sizes in ANOVA models for generalist teachers on CLASS

	Year One		Year Two	
	CLASS		CLASS	
	PC2	SA	PC2	SA
Subject	0.000*** (0.164)	0.139	0.000*** (0.454)	0.322
Grade	0.718	0.998	0.363	0.315
Sub*Gr	0.918	0.816	0.585	0.544
Subject	0.000*** (0.773)	0.610	0.000*** (0.790)	0.054
District	0.016* (0.028)	0.001*** (0.044)	0.014* (0.040)	0.001*** (0.057)
Sub*Dist	0.021* (0.026)	0.000*** (0.049)	0.054	0.040* (0.032)

Note: PC stands for principal component; SA stands for simple average. *** means the effect is significant at the 0.001 level. * means the effect is significant at the 0.05 level.

Table A 137: ANOVA with repeated measure on CLASS for Year One generalist teachers: The influence of grade level and district

GRADE

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size $\boldsymbol{\eta^2}$
ANOVA 1: CLASS Organization as dependent variable						
A: SUBJECT	66.933	1	66.933	85.221	0.000***	0.164
B: GRADE	4.228	3	1.409	0.566	0.718	0.003
A × B	0.394	3	0.131	0.167	0.918	0.001
Error(Within)	342.435	436	0.785			
Error(Between)	1367.261	436	3.136			
ANOVA 2: CLASS Simple Average as dependent variable						
A: SUBJECT	0.168	1	0.168	2.193	0.139	0.005
B: GRADE	0.009	3	0.003	0.012	0.998	0.000
A × B	0.072	3	0.024	0.313	0.816	0.002
Error(Within)	33.420	436	0.077			
Error(Between)	115.196	436	0.264			

DISTRICT

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
ANOVA 3: CLASS Organization as dependent variable						
A: SUBJECT	1135.997	1	1135.997	1325.145	0.000***	0.773
B: DISTRICT	38.081	4	9.520	3.080	0.016*	0.028
A × B	8.945	4	2.236	2.913	0.021*	0.026
Error(Within)	333.884	435	0.768			
Error(Between)	1331.690	435	3.061			
ANOVA 4:CLASS Simple Average as dependent variable						
A: SUBJECT	0.019	1	0.019	0.261	0.610	0.001
B: DISTRICT	5.029	4	1.257	4.964	0.001***	0.044
A × B	1.653	4	0.413	5.648	0.000***	0.049
Error(Within)	31.838	435	0.073			
Error(Between)	110.176	435	0.253			

Table A 138: ANOVA with repeated measure on CLASS for Year Two Generalist Teachers: The influence of grade level and district

GRADE

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
ANOVA 1: CLASS Organization as dependent variable						
A: SUBJECT	183.248	1	183.248	257.956	0.000***	0.454
B: GRADE	15.179	2	7.590	3.248	0.363	0.021
A × B	0.762	2	0.381	0.537	0.585	0.003
Error(Within)	220.219	310	0.710			
Error(Between)	724.447	310	2.337			
ANOVA 2: CLASS Simple Average as dependent variable						
A: SUBJECT	0.072	1	0.072	0.983	0.322	0.003
B: GRADE	0.476	2	0.238	1.161	0.315	0.007
A × B	0.089	2	0.045	0.610	0.544	0.004
Error(Within)	22.622	310	0.073			
Error(Between)	63.580	310	0.205			

DISTRICT

ANOVA 3: CLASS Organization component as dependent variable

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
A: SUBJECT	806.305	1	806.305	1158.205	0.000***	0.790
B: DISTRICT	29.286	4	7.322	3.175	0.014**	0.040
A × B	6.562	4	1.640	2.356	0.054	0.030
Error(Within)	214.420	308	0.696			
Error(Between)	710.340	308	2.306			
ANOVA 4: CLASS Simple Average as dependent variable						
A: SUBJECT	0.268	1	0.268	3.750	0.054	0.012
B: DISTRICT	3.631	4	0.908	4.627	0.001**	0.057
A × B	6.562	4	1.640	2.356	0.040*	0.032
Error(Within)	21.987	308	0.071			
Error(Between)	60.426	308	0.196			

APPENDIX J:

COMPARISON RESULTS AND ANOVA TABLES FOR MATHEMATICS TEACHERS

J.1. Summaries of Comparison and ANOVA Results

Table A 139: P-values and effect sizes for subject areas comparisons within mathematics

Instrument	Subject Areas	Comparison Level	Year One p-value (effect size)	Year Two p-value (effect size)
FFT	NO vs. AA (N = 230, 175)	Instruction	0.318	0.854
		Management	0.434	0.923
		Average	0.349	0.874
	NO vs. G (N = 135, 84)	Instruction	0.393	0.119
		Management	0.453	0.041* (0.227)
		Average	0.382	0.071
	AA vs. SP (N = 56)	Instruction	0.752	
		Management	0.671	
		Average	0.990	
CLASS	NO vs. AA (N = 231, 175)	Support	0.033* (0.141)	0.921
		Organization	0.719	0.151
		Average	0.046*(0.132)	0.861
	NO vs. G (N = 134, 84)	Support	0.399	0.991
		Organization	0.464	0.936
		Average	0.377	0.988
	AA vs. SP (N = 56)	Support	0.951	
		Organization	0.799	
		Average	0.934	
MQI	NO vs. AA (N = 221, 171)	Instruction	0.082	0.860
		Accuracy	0.811	0.633
		Average	0.157	0.974
	NO vs. G (N=125, 81)	Instruction	0.534	0.779
		Accuracy	0.005** (0.249)	0.282
		Average	0.030* (0.194)	0.812
	AA vs. SP (N = 55)	Instruction	0.899	
		Accuracy	0.032* (0.289)	
		Average	0.589	

Note: * means the difference is significant at the 0.05 level. ** means that the difference is significant at the 0.01 level. If not significant at least at the 0.05 level, only the p-value is provided.

Further examinations of other factors were performed on the above significant cases:

Table A 140: Summaries of P-values and effect sizes in ANOVA models for mathematics teachers

	NO vs. G			NO vs. AA		AA vs. SP
	FFT	MQI		CLASS		MQI
	Instruction p-value (effect size)	Accuracy p-value (effect size)	Average p-value (effect size)	Support p-value (effect size)	Average p-value (effect size)	Accuracy p-value (effect size)
Sub Areas	NA	0.486	0.529	0.220	0.322	0.102
Grade	NA	0.491	0.494	0.000*** (0.177)	0.315	0.086
Interaction	NA	0.643	0.523	0.429	0.544	0.884
Sub Areas	0.087	0.031* (0.038)	0.221	0.081	0.054	0.750
District	0.075	0.063	0.010** (0.105)	0.001*** (0.092)	0.001*** (0.057)	0.008** (0.244)
Interaction	0.721	0.395	0.325	0.997	0.040* (0.032)	0.152

Note: For Year One NO vs. G, 79 out of 84 teachers are in grade 4th, and only 1 or 2 teachers who are in other grade levels, so it is not analyzed to see if difference depends on grade level.

J.2. ANOVA Tables for Mathematics Teachers on FFT

Table A.140: ANOVA with repeated measure on FFT Management for Year Two mathematics teachers in Numbers & Operations and Geometry lessons

DISTRICT

ANOVA: FFT Management as dependent variable

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
A: SUB_AREA	4.271	1	4.271	3.011	0.087	0.037
B: DISTRICT	28.027	4	7.007	2.217	0.075	0.101
A × B	2.950	4	0.737	0.520	0.721	0.026
Error	112.043	79	1.418			
Total	249.700	79	3.161			

J.3. ANOVA Tables for Mathematics Teachers on CLASS

Table A 141: ANOVA with repeated measure on CLASS Support for Year One mathematics teachers in Numbers & Operations and Algebra & Algebraic Thinking lessons

GRADE

ANOVA 1: CLASS Support as dependent variable

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
A: SUB_AREA	26.388	1	26.388	1.510	0.220	0.007
B: GRADE	1590.361	3	530.120	16.310	0.000***	0.177
A × B	48.575	3	16.192	0.926	0.429	0.012
Error(Within)	3967.294	227	17.477			
Error(Between)	7377.950	227	35.502			

ANOVA 2: CLASS Simple Average as dependent variable

A: SUB_AREA	0.072	1	0.072	0.983	0.322	0.007
B: GRADE	0.476	2	0.238	1.161	0.315	0.007
A × B	0.089	2	0.045	0.610	0.544	0.012
Error(Within)	22.622	310	0.073			
Error(Between)	63.580	310	0.205			

DISTRICT

ANOVA 3: CLASS Support as dependent variable

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
A: SUB_AREA	54.662	1	54.662	3.067	0.081	0.013
B: DISTRICT	825.826	5	165.165	4.564	0.001***	0.092
A × B	5.966	5	1.193	0.067	0.997	0.001
Error(Within)	4009.903	225	17.822			
Error(Between)	8142.486	225	36.189			

ANOVA 4: CLASS Simple Average as dependent variable

A: SUB_AREA	0.268	1	0.268	3.750	0.054	0.012
B: DISTRICT	3.631	4	0.908	4.627	0.001***	0.057
A × B	0.724	4	0.181	2.536	0.040*	0.032
Error(Within)	21.987	308	0.071			
Error(Between)	60.426	308	0.196			

Note: *** means significant at the 0.001 level. * means significant at the 0.05 level.

J.4. ANOVA Tables for Mathematics Teachers on MQI

Table A 142: ANOVA with repeated measure on MQI Accuracy for Year One Mathematics Teachers in Numbers & Operations and Geometry

GRADE

ANOVA 1: MQI Accuracy ranks as dependent variable

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
A: SUB_AREA	0.031	1	0.031	0.489	0.486	0.004
B: Grade	0.050	1	0.050	0.478	0.491	0.004
A × B	0.041	1	0.041	0.643	0.424	0.005
Error(Within)	7.770	123	0.063			
Error(Between)	11.496	123	0.093			

ANOVA 2: MQI Simple Average ranks as dependent variable

A: SUB_AREA	0.028	1	0.028	0.398	0.529	0.003
B: Grade	0.046	4	0.046	0.471	0.494	0.004
A × B	0.036	4	0.036	0.523	0.523	0.004
Error(Within)	8.505	123	0.069			
Error(Between)	12.081	123	0.098			

DISTRICT

ANOVA 3: MQI Accuracy as dependent variable

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
A: SUB_AREA	0.549	1	0.549	4.742	0.031*	0.038
B: DISTRICT	1.408	4	0.352	2.299	0.063	0.071
A × B	0.477	4	0.119	1.030	0.395	0.033
Error(Within)	13.902	120	0.116			
Error(Between)	20.864	129				

ANOVA 4: MQI Simple Average as dependent variable

A: SUB_AREA	0.067	1	0.067	1.513	0.221	0.012
B: DISTRICT	0.723	4	0.181	3.511	0.010**	0.105
A × B	0.209	4	0.052	1.176	0.325	0.038
Error(Within)	5.341	120	0.045			
Error(Between)	20.864	129				

Table A 143: ANOVA with repeated measure on MQI Accuracy for Year One Mathematics Teachers in Statistics & Probability and Algebra & Algebraic Thinking

GRADE

ANOVA 1: MQI Accuracy as dependent variable

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
A: SUB_AREA	0.198	1	0.198	2.767	0.102	0.052
B: GRADE	0.966	2	0.483	2.573	0.086	0.093
A × B	0.018	2	0.009	0.124	0.884	0.009
Error(Within)	3.585	50	0.072			
Error(Between)	9.381	50	0.188			

DISTRICT

ANOVA 2: MQI Accuracy as dependent variable

Sources of Variation	*SS*	*df*	*MS*	*F*	P-value	Effect Size η^2
A: SUB_AREA	0.007	1	0.007	0.103	0.750	0.002
B: DISTRICT	2.523	4	0.631	3.870	0.008**	0.244
A × B	0.548	4	0.137	2.152	0.089	0.152
Error(Within)	3.055	48	0.064			
Error(Between)	6.232	48	0.130			

** means significant at the 0.01 level.

REFERENCES

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics, 25*, 95-135.

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. In R. Lissetz (Ed.), *Value Added Models in Education: Theory and Applications*. Maple Grove, MN.: JAM Press.

Bell, C. A., Drake, C., Wilson, M., Fraiser, A., & Kim, J. (2015). *Subjects-specific and general observation protocols as tools for the evaluation and improvement of teaching*. Paper presented at the The Annual Conference of the American Educational Research Association (AERA 2015), Chicago, IL.

Bill & Melinda Gates Foundation. (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study.

Boston, M., Bostic, J., Lesseig, K., & Sherman, M. (2015). A comparison of mathematics classroom observation protocols. *Mathematics Teacher Educator, 3*(2), 154-175.

Brennan, R. L. (2006). Perspectives on the evoluation and future of educational measurement. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 1-16). Westport, CT.: American Council on Education/Praeger.

Burgess, T. (2007). *Investigating the nature of teacher knowledge needed and used in teaching statistics.* (Unpublished doctoral dissertation), Massey University, Palmerston North, New Zealand.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757-783.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *American Mathemaical Monthly, 104*, 801-823.

Cohen, J. (2013). *Practices that cross disciplines?: A closer look at instruction in elementary math and English Language Arts.* (Doctoral of Philosophy), Stanford University.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Danielson, C. (2007). *Enhancing Professional Practice: A framework for teaching* (2 ed.): Association for Supervision & Curriculum Development.

Danielson, C. (2013). *The Framework for Teaching: Evaluation Instrument*. Princeton, NJ: The Danielson Group.

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement: A Review of State Policy Evidence. *Education Policy Analysis Archives, 8*(1).

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teahcer performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34*(2), 267-297.

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

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

Grossman, P., Cohen, J., & Brown, L. (2014). Understanding instructional quality in English Language Arts: Variations in PLATO Scores by Content and Context. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing Teacher Evaluation System: New Guidance from the Measures of Effective Teaching Project*. San Francisco: Jossey-Bass: A Wiley Brand.

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

Grossman, P. L. (1990). A study in contrast: Sources of pedagogical content knowledge for secondary English. *Journal of Teacher Education, 40*(5), 24-31.

Grossman, P. L., & Stodolsky, S. S. (1995). Content as context: The role of school subjects in secondary school teaching. *Educational Researcher, 24*(5), 5-23.

Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education, 38*(5), 427-437.

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 4,000 US early childhood and elementary classrooms. Retrieved from: http://fcdus.org/BuildingAScienceOfClassroomsPiantaHamre.pdf

Headden, S. (2011). Inside IMPACT: D.C.'s Model Teacher Evaluation System. Washington, D.C.: Education Sector.

Herbst, P., & Kosko, K. (2014). Mathematical knowledge for teaching and its specificity to high school geometry instruction. In J.-J. Lo, K. R. Leatham, & L. R. Van Zoest (Eds.), *Research Trends in Mathematics Teacher Education* (pp. 23-45): Springer.

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

Hill, h. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56-64.

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

Ho, A. D., & Kane, T. J. (2013). The reliability of classroom observations by school personnel: Bill and Melinda Gates Foundation.

Hull, J. (2013). Trends in Teacher Evaluation: How states are measuring teacher performance. Alexandria, VA. : Center for Public Education.

Joe, J. N., McClellan, C. A., & Holtzman, S. L. (2014). Reliability and the length and focus of classroom observations. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing Teacher Evaluation System: New Guidance from the Measures of Effective Teaching Project*. San Francisco: Jossey-Bass: A Wiley Brand.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4 ed., pp. 17-64). Washington, DC: American Council on Education/Praeger.

Kane, M. (2012). All validity is construct validity. Or is it? *Measurement:Interdisciplinary Research and Perspectives, 10*(1-2), 66-70.

Kane, M. (2013). Validting the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kane, T. J., Staiger, D. O., McCaffrey, D. F., Cantrell, S., Archer, J., Buhayar, S., & Parker, D. (2012). Gathering feedback for teaching: Combining high-quality observaitons with student surveys and achievement gains. Seattle, WA: Bill & Melinda Gates Foundation, Measures of Effective Teaching Project.

Kennedy, M. (2010). Attribution Error and the Quest for Teacher Quality. *Educational Researcher, 39*(8), 591-598.

Kennedy, M. M. (2004). *Examining teacher quality.* Paper presented at the Proceedings of the NCTM Research Catalyst Conference, Washington DC.

Kennedy, M. M. (2008). Sorting out teacher quality. *Phi Delta Kappan, 90*(1), 59-63.

Lipsey, M. W., Puzio, K., Yun, C., Herbert, M., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms: Institute of Education Sciences.

McCrory, R., Floden, R., Ferrini-Mundy, J., Reckase, M. D., & Senk, S. L. (2012). Knowledge of algebra for teaching: A framework of knowledge and practices. *Journal for Research in Mathematics Education, 43*(5), 584-615.

Meyer, J. P., Cash, A. H., & Mashburn, A. (2011). Occasions and the reliability of classroom observations: Alternative conceptionalizations and methods of analysis. *Educational Assessment, 16*(4), 227-243.

Michigan Council for Educator Effectiveness. (2013). Building an improvement-focused system of educator evaluation in Michigan: Final recommendations: Michigan Council for Educator Effectiveness.

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 System: New Guidance from the Measures of Effective Teaching Project*. San Francisco: Jossey-Bass: A Wiley Brand.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). A composite estimator of effective teaching. 51. Retrieved from: http://k12education.gatesfoundation.org/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

Milanowski, A. T. (2011). *Validity research on teacher evaluation systems based on the framework for teaching*. Paper presented at the American Education Research Association annual meeting, New Orleans, LA.

Moore, D. S. (1992). Teaching statistics as a respectable subject *Statistics for the twenty-first century* (Vol. 26, pp. 14-25). Washington, DC: Mathematical Association of America.

OECD. (2009). Evaluating and rewarding the quality of teachers: International practices: Organisation for Economic Co-operation and Development.

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 System: New Guidance from the Measures of Effective Teaching Project*. San Francisco: Jossey-Bass: A Wiley Brand.

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

Pivovarova, M., & Amrein-Beardsley, A. (2015). *Student growth percentiles (SGPs): Testing for validity and reliability.* Paper presented at the The Annual Conference of the American Educational Research Association (AERA 2015), Chicago, IL.

Rice, J. K. (2003). *Teacher Quality: Understanding the Effectiveness of Teacher Attributes*. Washington, D.C.: Economic policy Institute.

Rivkin, S. G., Hanushek, E. A., & Kain, J. E. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Salloum, S. J., Bieda, K. N., Sweeny, S. P., Torphy, K. T., Hu, S., & Lane, J. (2016). Capturing early career teachers' enactment of mathematics practices at scale. *Manuscript submitted for publication.*

Sanchez, E., & Blancarte, A. (2008). *Statistical thinking as a fundamental topic in training the teachers.* Paper presented at the Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education, Monterrey, Mexico.

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *Mathematics Education, 48*(1-2), 29-40.

Staiger, D. O., & Kane, T. J. (2014). Making decisions with imprecise performance measures: The relationship between annual student achievement gains and a teacher's career value added. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing Teacher Evaluation System: New Guidance from the Measures of Effective Teaching Project*. San Francisco: Jossey-Bass: A Wiley Brand.

Stodolsky, S. S., & Grossman, P. L. (1995). The impact of subject matter on curricular activity: An analysis of five academic subjects. *American Educational Research Journal, 32*(2), 227-249.

Strong, M. (2008). *Effective teacher induction and mentoring: Assessing the evidence*: Teacher College Press.

Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley Publishing Company.

Walkington, C., & Marder, M. (2014). Classroom observation and value-added models give complementary information about quality of mathematics teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing Teacher Evaluation System: New Guidance from the Measures of Effective Teaching Project*. San Francisco: Jossey-Bass: A Wiley Brand.

Wenglinsky, H. (2000). How teaching matters: Bringing the classroom back into discussions of teacher quality.