Generalizability of Scores from Classroom Observation Instruments

by

Mark C. White

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Educational Studies)
in The University of Michigan
2017

Doctoral Committee:

Professor Brian P. Rowan, Chair
Associate Professor Ben B. Hansen
Professor Brian A. Jacob,
Assistant Professor Matthew S. Ronfeldt

Mark C. White
mrkwht@umich.edu
ORCID ID: 0000-0003-2394-3151

Acknowledgements

I would like thank my dissertation committee for the thoughtful guidance and feedback they provided on my work. This dissertation would not be possible without them. I especially would like to thank my advisor Brian Rowan, not only for the countless hours he spent reviewing and providing feedback on this dissertation, but for the many opportunities he provided me as a research assistant and teaching assistant throughout my doctoral career. I've learned so much from him. I also want to thank Lesli Scott and Jennifer Smith for the guidance and support they have provided me over the years. Last, I want to thank my friends and family for all the support they give me.

Table Of Contents

List of Tables

List of Figures

List of Appendices

Abstract

This dissertation examined the effect of contextual features of the classroom environment on measures of teacher quality derived from classroom observation instruments. Using data on 228 teachers observed four times as part of the Understanding Teacher Quality (UTQ) study, scores from the Classroom Assessment Scoring System, the Framework for Teaching, and the Protocol for Language Arts Teaching Observation were analyzed using Generalizability Theory (GTheory) statistical models. The goals of these analyses were to examine how the reliability, validity, and bias of teacher quality estimates shifted as GTheory models successively incorporated a variety of contextual features of measurement present across specific occasions of observation. The contextual features of measurement examined included: (1) observation system design (SD) variables such as time of year and methods of scoring; (2) variables measuring curricular and instructional (CI) practices, such as whether a lesson focused on reading or writing or included discussion or lecture; and (3) variables measuring features of school organization (SO) such as classroom student composition.

Through comparison of models that successively adjusted for SD, CI, and SO variables, it was found that for all three observation instruments, over 20% of the variance attributed to teacher effects in traditional GTheory statistical models was due to sampling error stemming from values of the SD and CI variables recorded for a teacher on a given occasion of measurement, implying that traditional GTheory approaches (that do not incorporate SD and CI variables) can result in positively biased estimates of the reliability of teacher quality scores. However, this dissertation also found that teacher quality scores adjusted for SD and CI variables were highly correlated to teacher quality scores from

GTheory models that did not adjust for SD and CI variables, in part because occasions of observation were selected at random during the UTQ data collection period.

Inclusion of SO variables into GTheory statistical models had more far-reaching consequences. To begin, SO variables (and especially student composition) explained ~40% of the variance in teacher quality estimates, changed point estimates considerably, and reduced the reliability of these estimates to very low levels under normal observational designs. However, the decision as to whether or not to include SO facets into GTheory statistical models necessarily involves assumptions about whether these differences in observed teaching quality are driven by teacher sorting (e.g. better teachers are teaching more advantaged students) or co-construction (e.g., more advantaged students co-construct instruction along with the teacher, making instruction of higher measured quality in more advantaged settings). If one assumes that co-construction drives the observed effects of SO variables on teacher quality estimates, then these variables should be included in a GTheory model, but such inclusion will make teacher quality estimates very unreliable. If, on the other hand, one assumes that teacher selection drives the statistical relationship between SO variables and teacher quality estimates, inclusion of SO variables into a GTheory model would bias teacher quality estimates by removing this selection effect.

Overall, the results presented in this dissertation highlight the subtle ways that SD, CI, and SO variables can affect teacher quality estimates and how these subtle differences can affect the reliability, validity, and bias of teacher quality measures derived from classroom observation instruments.

Chapter I. Introduction

Teaching is an inherently situated activity (J. J. Cohen & Goldhaber, 2016). Teachers teach specific content to specific students using a variety of instructional formats (like lecture, recitation, discussion, or seat-work). Moreover, instruction unfolds over time. New days can involve teaching different content (like reading or writing) or teaching for different instructional goals (like introducing new content, reviewing, engaging in independent practice). Teacher's instructional choices can be dependent on their students and the content being taught. Further, teachers enact instruction in specific schools, and schools can vary in the ways they are organized and staffed for instruction and how they allocate students to teachers. Much research shows that the situated nature of teaching has the potential to affect the nature of instruction a teacher provides-- to specific groups of students, on specific days of the year, in specific school settings (e.g. Stodolsky, 1984).

This thesis explores how the situated nature of teaching complicates our ability to use classroom observation data measuring **teaching quality** (gathered over a particular set of *situated* instances of teaching) to make inferences about an observed teacher's ability to provide high quality instruction across *many* possible situated acts of teaching, which I will call **teacher quality**. While any specific measurement of teaching quality is drawn from a set of situated acts, the inferences we wish to make about a teacher's ability to provide high quality instruction are often more broad. For example, a principal may wish to draw conclusions about a teacher's ability to teach the full range of students likely to be encountered in a school, but the principal may only have data from a single class (or a few classes) taught by the teacher. A superintendent might wish to draw conclusions about which teachers are the best teachers across all schools in the district while having data on each teacher's instruction as it occurred in a single school. Researchers are often interested in estimating a stable teacher quality to engage in high quality instruction across even more school and district contexts, but again, have data only on instruction provided across a single school, on a limited number of days, where teachers teach a limited number of curriculum topics to a limited range of students. Each of these cases represents a problem of generalizing from observed scores on a small number of occasions of measurement to broader constructs that capture a teacher's ability to mount high quality instruction across a range of pre-defined contexts. Given the situated nature of teaching, this process of generalizing scores across contexts can be far more complex than most studies have acknowledged.

The process of generalization can be divided into two steps. In a first step, a measurement procedure is developed and a set of "facets" of measurement over which a user wants to generalize scores from that instrument is specified in advance. In the measurement literature, a facet is defined as any dimension of the measurement situation that may affect observed scores on the measure of interest and across which generalization of the observed score is desired. In much measurement work, the facets of measurement defined in advance consist of properties of the measurement protocol used to obtain an observed score. So, for example, if an analyst is using a classroom observation instrument to measure teaching quality, he or she might want to understand the extent to which a teaching quality score derived from a limited number of days of classroom observation can be dependably generalized to the score a teacher would receive had all potential days of instruction been observed. Alternatively, an analyst who has chosen to use a subset of items from a given observation instrument might want to understand how dependably a score derived from that limited item set can be generalized to a situation in which other subsets of items were used. Concerns with measurement protocols are important to the measurement of teaching quality, and I will examine these concerns in some detail in this dissertation. However, facets of measurement related to the measurement protocol itself are just one of many aspects of measurement across which an analyst might want to generalize. For example, when classroom observation instruments are being used to measure a given teacher's ability to mount high quality instruction, analysts might want to generalize scores from a given measurement protocol to several other facets of the measurement context in which measurement occurred, including across schools where a teacher might teach, across the students a teacher might teach, across the content domains a teacher might be expected to teach (e.g. reading, writing, algebra, geometry....), and so on.

Importantly, one need not seek to generalize across all possible levels of each of these facets. In fact, another step in a measures development is to define (in advance) not only the facets of measurement across which one wants to generalize, but also the levels of these facets across which generalization is sought. This exercise—of defining the facets and levels of facets across which one wants to generalize—defines what is called in the measurement literature the "universe" of generalization. Consider, for example, how the "universe" of generalization might differ for two different decision makers—a superintendent and a principal—working in the same district and using scores from the same classroom observation instrument to make inferences about a specific middle school teacher's ability to mount high quality instruction. The superintendent might want to use the teacher's score from the observation instrument to make inferences about how well that teacher will teach across all middle schools in the district; the principal, on the other hand, might want to use the scores only as an indication of how well the teacher teaches in the school where she was observed (and he serves as principal). In this case, the "levels" of the school facet over which these different administrators want to generalize constitute the "universes" to which generalization is desired, and, as we have just seen, the two administrators have defined two different universes based on their intended use of the scores.

A second step in generalizing scores obtained from a measurement instrument to a "universe" involves decisions about how to sample observations of teaching quality from the desired universe to which one wants to generalize. Imagine, for example, that we have designed an observation system that allows us to observe a teacher across a representative (or random) range of values from a facet's universe. Here, the universe plays the role of defining the range of values of a facet that a teacher may be observed over. Sampling over a representative range of values for the defined universe is necessary for proper generalization. For example, with a universe defined as all middle schools in a state, we might observe the same teacher as she works in urban, suburban, and rural middle schools, as well as middle schools with low and high proportions of minority students. This would, arguably, give a measure of the teacher in a representative set of schools, supporting the generalization of estimates of teacher quality to *all* schools in a state. The goal of observing teachers across a range of instances of each facet is to demonstrate how much observed teaching quality varies across the facet, allowing an estimation of how much each facet contributes to observed scores. Knowing how much teaching quality varies across facets enables an estimate of how dependably (i.e. reliably) teacher quality estimates represent a teacher's ability across facets, as I will discuss later in this thesis.

The two-step process of generalization just discussed often involves many practical problems. In particular, it is usually *not* possible to observe teachers across the full range of facets over which generalization is desired. For example, we might want to generalize scores for a teacher across a range of schools, but teachers are rarely observed teaching in different schools, and so we in fact do not have a very strong evidentiary basis upon which to examine how well observed teaching quality may generalize across schools. In such cases, we must extrapolate scores across schools. Extrapolation, here, in effect, is a process of generalizing across facets where no data exists to support the generalization. Extrapolation, instead, relies on arguments and/or assumptions to support the inference that scores apply across a facet. For example, we could argue that the knowledge and skills that teachers need to provide high quality instruction are common across schools. Thus, a teacher who provides high quality instruction in one school will similarly provide high quality instruction in another school, implying that teacher scores as traditionally estimated from classroom observations can be generalized across schools. Alternatively, we could argue that schools have different cultures and curricula (among other features), that these differences among schools affect observed teaching quality, and that this effect is constant across teachers. In this case, the best teachers in one school will be the best teachers in all schools, suggesting that teacher scores should be centered around school means and that, after this centering, scores are generalizable across schools. A third argument might claim teacher scores will differ across schools solely because of differences in the student composition of the school. In this case, teacher scores can be equated across schools by adjusting them for the effects of the student composition, much like is done with value-added score estimates. These "equated" scores are then generalizable across schools.

A problem faced in practice is that arguments justifying extrapolation can result in very different estimates of any given teachers' teaching ability. Further, in many cases, there will be little evidentiary basis upon which to accept one or another argument as "correct," highlighting the challenge of extrapolating scores to an unobserved universe. Thus, extrapolation provides a weak basis from which to "generalize" across facets and, whenever possible, teachers should be observed across the universe of facets to which one wishes to generalize a score so that data-based generalization can occur.

Unfortunately, this careful process of specifying the universe of generalization, sampling to provide evidence for generalizing, and providing arguments to support explicit extrapolation is not what occurs in practice when classroom observation instruments are used to make inferences about teaching ability. Instead, users of observation instruments typically define all days of instruction over the course of a year as the universe of generalization, sample a random subset of days, and then generalize observed scores to the "average" day of instruction. While this approach uses the same logic of generalization as just discussed, the situated nature of teaching and the difficulty of extrapolating is ignored in the process. And this creates at least three problems of measurement discussed in this thesis.

The first problem is instrument bias. Classroom observation instruments use a fixed definition of teaching quality, which (instrument developers assume), on average, captures high quality teaching practices. However, it could be that scores on a particular observation instrument do not correspond to high quality instruction in all cases. For example, most observation instruments focus on high cognitive demand as a feature of high quality instruction (due to the importance of students developing these critical thinking skills and the difficulty in teaching them to students). However, there may be specific points in time, specific types of lessons, or specific kinds of academic content where drilling on basic skills (rather than focusing on more cognitively demanding tasks) is the most beneficial approach for students. In these cases, classroom observation instruments emphasizing high cognitive demand may poorly measure actual instructional quality. The result will be instrument bias in observed teaching quality, which can result in biased estimates of a teacher's general ability to mount high quality instruction. Taking an approach to generalization that recognizes the situated nature of teaching allows for an exploration of instrument biases, as I will show in this thesis.

A second problem concerns efficiency in estimation. Assume that the teacher construct we are generalizing to includes teaching quality in both reading and writing lessons. When sampling is done randomly across the year, without regard to the content domain being taught, some teachers, by chance, will be observed only teaching reading and not observed teaching writing. Estimated teacher quality for this set of teachers will therefore include only part of the construct of interest (i.e. only teacher quality in teaching reading and not writing). Assuming teacher quality varies across content domains, this results in a poor estimate of teacher quality for the subset of teachers observed only teaching reading. Further, the randomness of whether a teacher is observed teaching reading or writing contributes to teachers' observed scores, leading to inefficient estimates of these scores (i.e. the construct-irrelevant randomness of sampling is included as a part of teacher score estimates). If, recognizing the situated nature of teaching, we instead observed each teacher in both reading and writing lessons, the sampling of days of instruction will lead to more efficient estimation of teacher scores, though as I discuss later in this thesis, this does lead to some subtle challenges related to shifting interpretations of the meaning of teacher quality.

The third problem arises from the need to extrapolate across contexts. Often, little to no attention is paid to the role that context might have in constraining or supporting high quality instruction. Comparisons of teachers across contexts necessarily involve assumptions about the effects of contexts on teaching quality. The default assumption--usually implicit--is that no context effects on observed teaching quality exist. There are, though, calls for observed scores to be adjusted for the student composition of classrooms (e.g. Whitehurst, Chingos, & Lindquist, 2014) and, in a similar vein, attempts to isolate how the students assigned to a specific class might affect estimates of teacher quality (Steinberg & Garrett, 2016). These approaches, however, are fairly rare at this point and are isolated to the problem of a classroom's student composition rather than recognizing potential context effects more broadly. Further, these approaches do not address the implications for the measurement of teacher quality more broadly, as I will show in this thesis. A situated view of teaching recognizes the impact of student composition on observed teaching quality as just one of a broad range of facets over which extrapolation is necessary[1]. There is a need for more explicitness about when data-informed generalization is being done and when extrapolation is occurring, including better-developed arguments to support the extrapolation at hand (and any alternative, reasonable extrapolation arguments).

I.1. The Problem

In this dissertation, I explore how the situated nature of instruction can affect the bias, reliability, and validity of estimates of teacher quality based on classroom observations[2]. As in much previous research on teaching, the main approach to estimating teacher quality from teacher observation data in my dissertation involves using statistical methods derived from Generalizability Theory (GTheory) to estimate the effect that specific teachers have on observed teaching quality net of other factors that potentially affect the teacher's scores (Brennan, 2001). One piece of the problem here is identifying important facets of measurement, which can include properties of the situation in which observed teaching occurs (e.g., the content being taught, the instructional formats in use, or the students being taught) as well as various properties of the overall protocol used to measure teaching quality (e.g. the specific items on the observation instrument being used to rate teaching quality, the procedures by which days of instruction are selected for observation, or the consistency with which raters score similar instances of instruction across items, days, and teachers). The second piece of the problem is identifying the ways in which all of these facets—teachers,

[1] Note that I am effectively ignoring the possibility of observing teachers across sections that would enable generalizing across student composition. There is an empirical question here. How much of the variation across student characteristics occurs within-teachers, between-sections (allowing generalization) and how much occurs between-teachers (necessitating extrapolation)? This will likely vary quite a bit across study designs and samples, but it seems reasonable, given residential sorting and teacher sorting within-schools that much of the differences in student composition will occur between-schools and between-teachers.

[2] Bias here represents whether the estimate matches the population value of the estimate. Validity, on the other hand, captures whether the estimate represents some meaningful notion of teacher quality and is often operationalized by situating the measurement within a broader nomological network of constructs. Estimates can be biased and valid (though this seems more rare) or unbiased and not valid. See Kane's (2006) distinction between generalizing and extrapolation for a broader discussion of this distinction.

situations, and conditions of measurement—affect observed scores and the resulting implications of these effects for the reliability, bias, and validity of inferences about teacher quality one can make from a particular set of classroom observation data. The last part of the problem is clearly identifying the boundaries of generalization and the inferences necessary to extrapolate to different desired ways of defining teaching quality.

I.2. Approach

My dissertation proceeds in three steps. In a first step I use a GTheory-inspired statistical model to examine issues related to the reliability with which differences among teachers in teaching quality can be estimated. This follows the traditional approach, ignoring the situated nature of teaching. In this section of the thesis, I focus mostly on decomposing variance in observed scores into a teacher component (i.e. teacher quality, which I view as the "true" score to be estimated) and other components reflecting deviations of observed teaching quality scores from the teacher quality estimate in the model (where these deviations are viewed as "error" variance). The statistical model I estimate here differs from those in the literature in that I include items as an important source of error variance (rather than modeling mean scores across items), which leads to a richer exploration of error components in teacher observations. This analysis provides a starting point for additional exploration of how aspects of the instructional context affect observed teaching quality as well as a point of comparison for statistical models estimated at later points in this thesis.

In the second step of the dissertation, I expand the GTheory variance decomposition just discussed so that it now attends to effects that arise from contextual features of the lessons being observed (e.g. the content being taught, the instructional formats in use, the characteristics of students in the classroom, and the school where the teaching occurs). The inclusion of these additional "facets" of measurement in a GTheory analysis is not common in most reported research. The goal of this step is to understand how features of the teaching context affect observed teaching quality and to explore the implications of these effects for the reliability, bias, and validity of estimates of teacher quality. In this section of the dissertation, I present three additional statistical models that make corrections for various types of facets. The first model adds controls for score artifacts that arise from when and how scores are collected (e.g. date scoring occurred, whether scoring was done live or by video). These factors arguably should be controlled for in any estimate of teacher quality in order to reduce sampling variability in scores. The second statistical model adds controls for the content being taught and the instructional formats in use in the lessons observed. This model adjusts for inefficiencies (i.e. reduces sampling error) stemming from randomly sampling days by estimating a teacher effect that captures a teacher's ability to engage in a range of valued forms of instruction rather than the average instruction provided over the course of a year. This model also allows me to test for potential sources of instrument bias. The third statistical model adds controls for the contexts in which teachers teach (e.g. grade taught), moving towards the problem of extrapolating scores across contexts. The contribution of this section of the dissertation is not to advocate for one specific approach compared to another for arriving at estimates of the teacher quality construct. Rather, it is to show how carefully considering the situated nature of teaching and the construct one wishes to generalize to can lead to many approaches for score estimation, each of which has different implications when it comes to assessing the reliability, bias, and validity of resulting estimates.

A final step in the data analysis explores the validity of estimates of teacher quality derived from different statistical models, where a teacher’s value-added score (i.e. a measure of the average gains in learning experienced by students in a teacher’s class) is used as a criterion variable to examine concurrent validity of estimates of teacher quality. In this step, the goal is to understand whether it is possible to make claims about the relative validity of teacher quality measures derived from different models. As I will argue, the various models that I present can increase measurement precision and either reduce bias or increase bias, depending on the assumptions we are willing to make about the nature of teacher quality and the differences in teachers across school contexts. Exploring the validity issues across models provides more information about the construct being measured by each model. Note, however, that rather than searching for the "best model", viewed from the framework of this thesis, what the validity analysis really is intended to examine is which model provides an estimate of teaching quality that is most aligned with the definition of teaching quality implicit in value-added scores. That alignment, however, should not be taken as an ironclad rule about which model-based estimate is "valid." Instead, according to the arguments developed in this thesis, validity depends crucially on the features of context across which one wants to generalize, and it is explicit arguments about generalization that determine which (of many plausible) models should be used.

I.3. Data

In order to explore the research problems just discussed, I use data from the Understanding Teaching Quality (UTQ; http://utqstudy.org/) project, which gathered data on teaching quality from 228 English Language Arts (ELA) teachers in grades 6-8 using three, widely-used, classroom observation instruments: the Classroom Assessment and Scoring System (denoted as CLASS here and described in Pianta, Hamre, Haynes, Mintz, & La Paro, 2007); the Framework for Teaching (denoted as FFT here and described in Danielson, 2000); and the Protocol for Language Arts Teaching Observations (denoted as PLATO here and described by Grossman, Loeb, Cohen, & Wyckoff, 2013). The UTQ data is sufficient for me to estimate the effects of many different facets of measurement (including teachers, contexts of teaching, and conditions of measurement) on measured teaching quality. As I discuss in more detail in Chapter 4, each teacher in this study was observed on 4 days of instruction, spread across two separate class sections of students, by multiple raters recording scores on many different items from each of the classroom observation instruments. In addition, as part of the data collection regime, the content being taught on a given occasion of measurement was recorded, as was the type of instructional interactions taking place. Finally, the UTQ data set includes data allowing me to measure the school location of each teacher, the prior achievement levels, and socioeconomic characteristics of students in different class sections taught by that teacher, and the value-added scores of teachers for the study year and the year prior.

I.4. Outline of Dissertation

In the next chapter (Theoretical Framework), I provide a detailed overview of the theoretical framework used to frame this study. I start by providing an in-depth introduction to Generalizability Theory (GTheory), focusing on how it separates true score variance from error variance and, especially, how it isolates and describes sources of error variance, allowing a deeper understanding of how classroom observation instruments function as tools of measurement. I then discuss how contextual features of measurement can be incorporated into GTheory models, tying their incorporation to the problem of generalizing observed scores across contexts. Next, I discuss the various ways that contextual features of measurement might affect estimates of teacher quality. I conclude the second chapter by creating three categories of measurement facets that differ in how they are likely to affect the measurement process. Some contextual features will only make measurement more inefficient, some may bias estimates of teacher quality, and others may do both.

In the third chapter (Review), I present a literature review on past measurement work on observation instruments. I highlight previous uses of GTheory to study modern observation instruments. I will mostly focus on what is known about how contextual features of the measurement context affect observed teaching quality. Where possible I try to describe what is known about how these contextual features impact estimates of teacher quality, but past work has not focused much on these constructs.

The fourth chapter (Methods) starts by expanding upon the brief introduction to the data provided here. I then provide detailed introductions to the various statistical models that I will be using. I conclude the chapter by discussing my analytic approaches, including how I will identify instrument bias, how I explore the effect of measurement facets on observed teaching quality, and how I explore the validity of estimates of teacher quality. Underlying each of these analyses is the effect that accepting the situated nature of teaching has on how we think about generalizing observed scores to create estimates of teacher quality.

The fifth chapter (Results) presents results for the three approaches described here. I start by showing the relatively small role that teacher quality plays in explaining observed teaching quality. Next, I show that items and raters both play a large role as sources of error in observed teaching quality. I next highlight the large impacts that contextual features have on observed teaching quality across all categories of contextual features. This, however, does not produce a large impact on estimated teacher quality, likely due to the near random sampling of days and random assignment of raters in the UTQ study. Only the grade teachers teach and the characteristics of students in their classroom have a meaningful, though modest, impact on estimates of teacher quality. Especially important here is that models that do not account for the situated nature of teaching lead to imprecise estimates of teacher quality, which leads to inflated estimates of the reliability of scores. Last, I show that there is no evidence of differential validity of teacher quality across models, which seems to be the result of low power, given the small differences in teacher quality estimates across models.

In the last chapter (Discussion), I connect the results to their implications and the prior literature. I focus the discussion on the evidence that I found for the bias, reliability, and validity of scores from classroom observation instruments. I highlight the distinction in implications for both research and practice. Especially important for the effect on estimates of teacher quality is the careful sampling that is a hallmark of research, but less possible in practice. Throughout this chapter, I discuss limitations of this thesis and next steps for the research.

Chapter II. Theoretical Framework

In this chapter, I discuss the analytic and theoretical frameworks for this dissertation. As discussed in the preceding chapter, teaching is a situated activity, which makes estimating teacher quality from teaching quality complex. In this chapter, I introduce Generalizability Theory (GTheory) as a means for statistically modeling this complexity. I first describe the basic ideas underlying GTheory and show how the language of GTheory can be applied to the problem at hand. This discussion will show that GTheory allows me to model observed teaching quality scores as containing both a true score component and multiple error components. In the next section of the chapter, I show how the typical GTheory model requires the researcher to declare in advance what features of the measurement context are to be considered sources of errors in measurement (e.g. raters, items, days) and note that these features are called “facets” of measurement in GTheory. These defined facets, we shall see, help define the “universe” of measurement situations over which observed scores are intended to be generalized. Importantly, however, GTheory also recognizes that some features of a measurement context might not easily be taken into account or included explicitly as facets in the GTheory statistical model. In GTheory, these are called “hidden facets.” In the next section, I detail different ways that hidden facets might impact observed teaching quality and the implications for the reliability, bias, and validity of estimates of teacher quality. Next, I discuss three classes of “hidden facets” that I intend to analyze as part of this dissertation and discuss how these hidden facets might affect the inferences we can draw about teacher quality from a given set of classroom observation data. I conclude the chapter by reviewing the pertinence of the “hidden facets” problem to the research questions at hand.

II.1.Generalizability Theory

Generalizability Theory (GTheory; Brennan, 2001) forms the foundation of my analytic approach to making inferences about teacher quality from classroom observations. GTheory focuses on the problem of generalizing scores across facets of measurement and follows the same approach to generalizing that I described in the introduction. GTheory starts with a definition of the construct of interest. For classroom observation instruments, this is usually a teacher's ability to create high quality instruction. In the discussions that follow, I use the term teacher quality to represent the construct of interest in all models, though it must be noted that there are many subtly different ways of defining "teacher quality". Teacher quality can be defined as a teacher's ability to teach in a given school, the average of the observed teaching quality provided over the course of a year, a teacher's ability to teach a range of important curriculum topics, or a teacher's ability to teach across a range of school contexts. The distinction between these definitions is important (and detailed in later chapters) because the way teacher quality is defined determines the universe to which scores generalize, the appropriate model for estimating teacher quality, and the aspects of teaching quality that affect teacher quality. For example, defining teacher quality as school-specific precludes comparing teachers across schools while defining teacher quality as existing across schools requires some justification for extrapolating scores across contexts.

The next step for GTheory is to define the features of the measurement context (i.e. facets) across which observed teaching quality will vary and to which generalization is desired. As described before, this includes specifying the full range of "levels" a facet might take, which is called the facet's domain. The first thing to note about this process is that the facets of measurement depend on the specific definition of teacher quality. If the construct is a teacher's ability to teach within his or her school, school is not a facet because the teacher quality construct is defined as referring only to quality in the teacher's current school.

Importantly, because school is not a facet (and the construct is defined within schools), comparisons of teacher scores cannot be made across schools. If the construct is a teacher's ability to teach in any urban school in the state, school is a facet because a teacher's observed teaching quality likely varies across schools. The facet's domain defines the universe to which the school facet generalizes, in this case, urban schools within the given state.

Part of the definition of a construct, then, involves defining the contexts across which generalization is desired, which, in turn, defines the facets of interest. Of course, numerous facets exist, for example: classroom student composition, content domain being taught, time of year, time of day, the rater doing the scoring, and innumerable other facets. While there is a need to recognize the situated nature of teaching by recognizing features of the classroom and day context across which observed teaching quality varies, all facets cannot possibly be included and modeled. GTheory models generally are based around a set of facets that are explicitly planned to vary through the design of the observation system. Note that, in fact, the teacher is also facet of measurement and treated conceptually like any other facet, though, of course, the teacher, being the focus of measurement typically gets more attention than other facets. The typical facets characterizing the planned variation are teachers, items, raters, occasions of measurement, days, and sections (e.g. Kane et al., 2012; Mashburn, Downer, Rivers, Brackett, & Martinez, 2013). These planned facets broadly characterize some of the situations over which observed teaching quality is measured, but they do not capture the situated nature of teaching fully. Those features of context not measured are called hidden facets because they are “hidden” from the estimation model. Note that with this conceptualization, the distinction between hidden and non-hidden facets is model and protocol dependent. In this thesis, I am interested in studying some of these hidden facets directly.

Specifying the construct creates the framework of GTheory, but this framework needs to be reflected in the data. The data comes from sampling teachers' instruction. For each facet to be generalized across, a teacher needs to be observed across a representative sample of levels from the facet's domain. GTheory generally assumes both that sampling occurs randomly (or ignorably) from all possible levels of facet and that each level of a facet is an equally good substitute for any other level of that facet. For example, taking the day facet, it assumes that every day (of instruction) is equally likely to be observed and that there is no reason to prefer sampling day 1 as compared to day 2 or 9. This sampling facilitates generalization because teaching quality is observed across a representative set of levels from a facet's domain, allowing a prediction of how teaching quality will look across unobserved levels from that facet. That is, by analyzing how observed teaching varies across sampled levels of a facet, we can predict how scores will vary across all levels of the facet and so how stable scores will be across the facet. This is the theoretical basis that justifies estimating the reliability of scores in GTheory. As I will discuss below, however, the complexity stemming from the situated nature of teaching creates challenges for this framework. Namely, each day of instruction is not necessarily an equal representation of teacher quality.

This process of specifying the construct and sampling across facets in a way to allow generalization is the foundation on which GTheory statistical models are built. Similar to classical test theory, GTheory models observed scores (X) as being composed of a true score component (T) plus an error (ϵ) component (i.e. $X = T + \epsilon$). The true score here represents the previously defined construct of interest (i.e. teacher quality). The error term in GTheory (ϵ) is further decomposed as the sum of independent contributions from each planned facet and the interactions of all facets. For example, if we assume that the only facets of measurement are raters and days, the error term will be broken down into a component caused by raters, a component caused by days, a component caused by the interaction of raters and days, and a residual error (i.e. $\epsilon = \upsilon_{raters} + \upsilon_{days} + \upsilon_{raters-by-days} + \upsilon_{residual}$). Each of the facets included as part of the error term are called error facets and these error facets are assumed independent[3].

Of more interest in GTheory models, usually, is the relationship between the variances of the modeled terms (i.e. $var(X) = var(T) + var(\upsilon_{raters}) + var(\upsilon_{days}) + var(\upsilon_{raters-by-days}) + var(\upsilon_{residual})$). This is because the variances determine the relative importance of the true score and various error facets and the reliability of score estimates. When the variance of the true score is large (relative to the error facets), the observed scores measure teacher quality well. Error facets with large variances represent the largest sources of error in observed scores. These large facets can then be the focus of targeted efforts to improve the functioning of observation scores as a measure of teacher quality. For example, if the rater error facet is large, more effective training for raters should help improve measurement.

The reliability of the estimate of teacher quality (i.e. T or υ_T below) also comes directly from the estimated variances. Reliability is defined as the percentage of the observed score variance that is due to the true score (i.e. $var(T)/var(X)$). This is easily calculated with the variances from the GTheory model. In fact, GTheory is often used to conduct what is called a "decision study", which examines the impact on the reliability of score estimates of averaging teacher scores across multiple measurements. Consider what happens when using the average score from two separate raters who independently scored an occasion of instruction as an estimate of teacher quality. The same true score contributes to the scores given by each rater, but the rater effects (i.e. deviations from true score caused by raters) are different and independent. The variance of the average of two independent, identically

[3] In fact, when data are balanced across facets and all interactions are modeled, the error facet become independent by design (i.e. the assumption must be true).

distributed random variables is half of the variance of the original variable (i.e. the variance of the rater error facet is cut in half when averaging scores across two independent raters). Thus, the variance of the true score remains constant while the variance of the error facets (namely the rater facet) is cut in half. Averaging scores across levels of a facet increases the reliability of estimates of teacher quality in a calculable way. Decision studies, then, use this fact to estimate score reliabilities across different sampling designs. For example, a decision study would estimate the score reliability stemming from averaging scores across four days with one rater each day; averaging scores across 3 days with two raters each day; and, more broadly, 'A' days with 'B' raters each day. In this way, GTheory analyses can flow directly into examining the reliability of score estimates, including predicting score reliability for future observational system designs.

In summary, GTheory provides a theoretical and analytic framework from which to generalize observation scores across a number of different facets through emphasizing clarity in defining the construct of interest and facets across which generalization is desired. In doing so, GTheory helps to show which facets of measurement have the largest impact on observed teaching quality. It also provides a framework for understanding the reliability of teacher quality estimates and estimates a true score, which represents teacher quality. It does not, however, directly address the issue of extrapolating teacher scores across facets where no data exists to support generalization.

II.1.1. A Full GTheory Model for Classroom Observation Data Up to this point, I've focused broadly on how GTheory is used and why it is relevant for exploring the generalizability of classroom observation instruments. In this section, I provide a full GTheory measurement model for typical classroom observational data (such as the UTQ data that I use in this thesis). I present this model to demonstrate the complexity of GTheory models and to provide a basis for further discussions about the specific impacts of characteristics of the measurement context on observed teaching quality.

In UTQ, observation instruments were used to measure teaching quality for a given teacher using multiple items across multiple days of instruction in two sections, where scores were given by multiple raters. Further, each day of observation was scored as multiple occasions, which were created by dividing days into 15 minute intervals. Thus, using GTheory, a goal is to assess the generalizability of observed teaching quality across all potential levels of items (I), raters (R), occasions of measurement within days (O), days of instruction within sections (D), and sections within teachers (S). The approach described here follows that of much previous research, though it uses a more complex (i.e. complete) model (e.g. Bell et al., 2012; Casabianca et al., 2013; Hill, Charalambous, & Kraft, 2012a; Ho & Kane, 2013; Kane et al., 2012; Mashburn et al., 2013). Effects stemming from the interactions of facets are also modeled as sources of variation. For example, the rater-by-day effect captures whether, after controlling for day and rater main effects, a rater is more lenient than expected on a specific day due to an idiosyncratic reaction to the day scored. The measurement model just described would be denoted as $i \cdot r \cdot (o\colon d\colon s\colon t)$ in GTheory (i.e. as occasions nested within days nested within sections nested within teachers crossed with items and raters). This gives four levels of nesting (occasions, days, sections, and teachers) and a total of 19 facets that affect observed scores. Most work (including this dissertation) does not fully model all facets of this full measurement model. For example, some researchers choose to average across items before conducting GTheory analysis, leading the analysis to focus only on average scores and reducing the model to 9 error facets. Written as a statistical model, the model assumes that observed scores, $X_{ir(o:d:s:t)}$, vary around an overall sample mean with a deviation from this mean due to each of the 19 facets:

$$
\begin{aligned}
\mathrm{X}_{\{ir(o:d:s:t)\}} = \mu + \upsilon_{t} &+ \upsilon_{\{s:t\}} + \upsilon_{\{d:s:t\}} + \upsilon_{\{o:d:s:t\}} \\
+\upsilon_{i} &+ \upsilon_{\{it\}} + \upsilon_{\{i(s:t)\}} + \upsilon_{\{i(d:s:t)\}} + \upsilon_{\{i(o:d:s:t)\}} \\
+\upsilon_{r} &+ \upsilon_{\{rt\}} + \upsilon_{\{r(s:t)\}} + \upsilon_{\{r(d:s:t)\}} + \upsilon_{\{r(o:d:s:t)\}} \\
+\upsilon_{\{ir\}} &+ \upsilon_{\{irt\}} + \upsilon_{\{ir(s:t)\}} + \upsilon_{\{ir(d:s:t)\}} + \upsilon_{\{ir(o:d:s:t)\}}
\end{aligned} \tag{1}
$$

where o is occasions, d is days, s is sections, t is teachers, i is items, r is raters, μ is the overall average quality and υ_{xy} generally refers to deviations from this mean resulting from unique combinations of facets x and y. $x{:}y$ denotes that facet x is nested in facet y. The variance of facet x is $var(\upsilon_x)$. $\upsilon_{s:t}$ is section deviation from teacher quality, $\upsilon_{d:s:t}$ is day deviation from average section quality, $\upsilon_{o:d:s:t}$ is occasion deviation from average day quality[4]; υ_i represents item difficulty or centered average item scores; and υ_r is a rater leniency effect. υ_{it}, $\upsilon_{i(s:t)}$, $\upsilon_{i(d:s:t)}$, and $\upsilon_{i(o:d:s:t)}$ model teacher, section, day, and occasion quality varying across items (i.e. difficulty of items varies across teacher/section/day/occasion); υ_{rt}, $\upsilon_{r(s:t)}$, $\upsilon_{r(d:s:t)}$, and $\upsilon_{r(o:d:s:t)}$ model separate raters ranking teachers/sections/days/occasions differently (e.g. rater bias or halo effects); υ_{ir} represents raters differing on an item's difficulty; υ_{irt}, $\upsilon_{ir(s:t)}$, $\upsilon_{ir(d:s:t)}$, and $\upsilon_{ir(d:s:t)}$ allow raters' scores of teacher/section/lesson/occasion quality to vary across items (e.g. rater unreliability or item specific rater bias).

II.2.Hidden Facets

In this section, I start to consider the impact of the situated nature of teaching on observed teaching quality, moving beyond those facets that are a planned part of measurement protocols. The fact that important facets are often not part of the planned

[4] I assume here occasions are independent and nested within days. This is incorrect in that occasions are ordered in time. The first occasion within a lesson is unique in a specific way, as is the second, third and so on. This suggests that occasions could be crossed rather than nested. This would estimate a unique effect for the first occasion as distinct from the effect of the second or third. The interaction of occasion and day would then capture what I am calling occasion. This would increase the complexity of the model by adding additional facets (an occasion order main effect and up to nine interaction terms). The current model is simpler and better captures the structure of accountability systems where informal observations can be conducted over any 15 minute occasion. In later models, I will account for this structure of occasions through fixed effect moderators.

features of measurement is recognized by GTheorists. Thus, GTheory enables an exploration of the import of facets that characterize the situated nature of teaching (e.g. content domain taught, students taught, school context...). These facets are, in the language of GTheory, *hidden facets* because they are measurement facets (i.e. sources of score variation) not explicitly modeled in Equation (1). When hidden facets are not modeled, the variance in observed scores due to the teacher (i.e. the true score variance) or any other error facet may be either over- or under-estimated (Webb, Shavelson, & Haertel, 2006). For example, assuming that writing and reading lessons receive different scores on average, two otherwise equivalent teachers may still receive different teacher quality estimates if one happens, by chance, to be observed teaching writing (rather than reading) more than the other teacher. Since all days are not equal representations of the underlying construct of teacher quality, the hidden facet, through the random sampling of days, can contribute additional variation to teacher scores (i.e. sampling error gets included in the teacher quality estimate, $\hat{v_t}$). The result is an inflated estimate of the variance in teacher scores (i.e. $E(\hat{var}(v_t)) > var(v_t)$), which is only one of a number of possible effects of hidden facets. The exact effect of the hidden facet will depend on whether it acts within-teachers or between-teachers and on the distribution of the hidden facet across teachers.

Before discussing the possible effects of hidden facets on observed teaching quality in detail, I must define some key terms necessary for this discussion. First, true teaching quality on a given occasion of measurement can be denoted as $X_{ir(o:d:s:t)}^{true}$ and is the extent to which effective instructional interactions occur in the classroom on an occasion of measurement, where effective interactions are ones that promote student learning. This may vary from the observed teaching quality defined by Equation (1) (i.e. $X_{ir(o:d:s:t)}$), where any difference between these two values represents bias in observed teaching quality. Second, true teacher quality is symbolized by v_t in Equation (1) and is simply the effect of a teacher on observed teaching quality as estimated across numerous occasions of measurement. Note again that the precise meaning of υ_t is model dependent, but as it appears in Equation (1), true teacher quality is the average of the observed teaching quality provided across the full observation period (i.e. from the first occasion that could be sampled through the last that could be sampled). υ_t cannot be directly known, but we can use models to estimate teacher quality (υ_t) and the estimate is denoted as $\hat{\upsilon}_t$ and termed estimated teacher quality. Now, consider what happens if we add to Equation (1) a "hidden" facet called facet-a. For simplicity, I will assume facet-a is positively related to observed teaching quality, and takes on values 0 (a_0) and 1 (a_1). Facet-a could, for example, be writing instruction where facet-a_0 indicates no writing instruction and facet-a_1 indicates writing instruction took place. Adding this hidden facet to Equation (1) results in an equation similar to Equation (1), but with a fixed effect, β_a, added (i.e. μ is replaced with $\mu + \beta_a$). The parameters of this equation are written with an 'a' superscript (e.g. υ_t^a). I call this the adjusted equation because it "adjusts for" the main effect of facet-a, predicting higher quality teaching when observing facet-a_1 than when observing facet-a_0. Note that β_a is simply a regression coefficient for dichotomous variable facet-a. This removes the average effect of facet-a from estimates of teacher quality ($\hat{\upsilon}_t$). In what follows, I focus on the impact that the hidden facet has on estimated teacher quality (i.e. comparing $\hat{\upsilon}_t$ and $\hat{\upsilon}_t^a$) and the variances of teacher quality (i.e. $\widehat{var}(\upsilon_t)$ and $\widehat{var}(\upsilon_t^a)$).

The switch from the unadjusted model (Equation 1) to the adjusted model (which includes the β_a effect) changes the interpretation of the teacher quality estimate, not just the value of the estimate (i.e. υ_t and υ_t^a have slightly different meanings). This shift in meaning is best explored by examining which aspects of teacher quality the unadjusted and adjusted models include in their estimates of teacher quality. The unadjusted model estimates teacher quality without regard to facet-a. Teacher quality (i.e. υ_t) in this model reflects the average quality of instruction provided over the observation period. Part of teacher quality in this model is the frequency with which teachers engaged in instruction at each level of facet-a. This reflects the belief that the decision of what to teach has important effects on what students learn and thus should be thought of as an aspect of teacher quality (e.g. Polikoff & Porter, 2014). This decision of what to teach is included in teacher quality (i.e. υ_t), which, as I have argued, leads to an additional source of sampling error, the frequency with which a teacher is observed teaching at each level of facet-a (which can also be understood as error in the prediction of how often a teacher engages in facet-a_0 and facet-a_1 instruction from how often teachers are observed in engaging in instruction at each level of the facet).

The adjusted model "adjusts for" the fact that facet-a_1 lessons, on average, score β_a points higher than facet-a_0 lessons. Teacher quality (i.e. υ_t^a) in this model now reflects an average of the quality of instruction provided in facet-a_1 and the quality of instruction provided in facet-a_0 over the observation period. The adjusted model "equates" instruction across the levels of facet-a so differences in the frequency with which facet-a_1 and facet-a_0 are taught is no longer included in teacher quality (i.e. υ_t^a purposefully excludes the teacher's choice to teach facet-a_1 lessons versus facet-a_0 lessons). This also removes the sampling error coming from the frequency with which teachers are observed across levels of facet-a, under the assumption that the β_a parameter correctly models the facet's effect on observed teaching quality.

Thus, the difference in teacher quality between models is whether the frequency that teachers teach facet-a_1 and facet-a_0 is part of teacher quality. However, even under the assumption that the frequency teachers teach facet-a_1 and facet-a_0 is part of teacher quality, there may be reason to prefer the adjusted model. This is because the gain in the precision of the teacher quality estimate in the adjusted model (over the unadjusted model) can outweigh the bias that arises in the adjusted model (due to adjusting away part of true teacher quality). That is, there is a bias-variance tradeoff. If we cannot reliably estimate the frequency with which teachers teach facet-a_1 and facet-a_0 during the observation period using the observed occasions, the reduced sampling error of the adjusted model is likely preferable because, without being able to estimate this frequency, we cannot estimate how differences in the same frequency may affect teacher quality. Prior research suggests that up to 15-30 days are needed to accurately estimate the frequency with which teachers engage in instruction across content domains (Rowan, Camburn, & Correnti, 2004). However, observation protocols generally sample about four days of instruction per teacher. Thus, it seems that the gain in precision from the adjusted model may often outweigh the possible introduction of bias. This trade-off between adjusting for facet-a to eliminate the sampling error associated with facet-a and introducing bias by adjusting for facet-a varies based on the observation protocol, the meaning of facet-a, and one's belief about what should and should not contribute to teacher quality. The question of what should and should not contribute to teacher quality is key to this thesis. In the next section, I discuss the ways that hidden facets affect teaching quality and the resulting impact this can have for estimates of teacher quality.

II.2.1. Average Differences in Teaching Quality across Levels of Hidden Facets

The simplest empirical test for whether facet-a affects observed teaching quality ($X_{ir(o:d:s:t)}$) is to test whether β_a is significantly related to observed teaching quality (i.e. test if the regression coefficient $\beta_a$is non-zero). A statistically significant effect implies there is something unique about facet-a_1 that affects observed scores. The effect could be significant for one of two possible reasons. First, the instrument may be biased for days with facet-a_1 (i.e. $E[X_{ir(o:d:s:t)}] \neq X_{ir(o:d:s:t)}^{true}$ for $a = 1$). Bias occurs when factors unrelated to teaching quality affect the observed score. For example, an instrument focused on classroom interactions might be positively biased for classroom discussions, rating all discussions higher than their true teaching quality would warrant. Instrument bias on observed teaching quality will generally bias estimates of teacher quality. The adjusted model should correct for this instrument bias, assuming the bias manifests as an average difference in observed teaching quality across levels of facet-a. This correction occurs because the difference in average scores on a given level of facet-a is removed by the adjustment of β_a. In the empirical analyses discussed later, I will capitalize on the fact that UTQ includes three separate observation instruments to test for potential instrument bias, assuming that any differences in the magnitude of β_a across the three instruments are indicative of bias. For example, if one instrument finds a positive effect of β_a while the other two find a negative effect of β_a, bias must exist. In this case, true teaching quality is either higher, lower, or the same for facet-a_1 (as compared to facet-a_0), but it cannot be both higher and lower at the same time. Of course, I am unable to determine which of the instruments is biased in a given case since I have no direct measure of true teaching quality. All I know is that one instrument or another is biased. This is unfortunate because, in the presence of bias, the models adjusted for facet-a should provide a better estimate of teacher quality (i.e. υ_t^a is more valid than υ_t) so knowing which instruments show bias and should estimate teacher quality with an "adjusted" model would be useful. Note that all three instruments used in the UTQ study emphasize interactive forms of instruction and higher-order thinking skills, which limits my ability to examine instrument bias since all three instruments may share similar biases.

The second reason for a mean difference in observed scores across levels of facet-a is that true observed teaching quality varies across levels of facet-a. When this occurs, the impact of the hidden facet on the parameters of the measurement model will depend on whether the hidden facet is within-teachers or between-teachers. Within-teacher hidden facets are sampled within-teachers across the full range of the facet's domain. This means that data exists to support generalizing across these facets, at least for many teachers. Between-teacher facets take on only a single level (or a very limited subset of levels) for a given teacher. Thus, generalization is impossible and we must extrapolate to equate scores across these facets. I discuss these two types of hidden facets separately below.

II.2.2. Within-Teacher Hidden Facets Within-teacher hidden facets occur when the average teacher displays higher teaching quality when observed with facet-a_1 than when observed with facet-a_0[5]. Because this variation is occurring within teachers, under most definitions of teacher quality, which usually assume a stable construct across days (and often sections), this variation in scores is a source of error. Assuming that facet-a varies across days of instruction, the random sampling of days will lead to random variation in how many days teachers are observed at each level of facet-a. This random variation leads to variation in teacher scores that is unrelated to teacher quality. Thus, when there is an average difference in scores across levels of facet-a, within-teacher hidden facets will lead to estimates of the variance of teacher quality that are positively biased (i.e. $E[\hat{var}(v_t)] > var(v_t\ )$). Models that adjust for the effect of facet-a provide a better estimate of day variance (i.e. $E[\hat{var}(v_t^a)] = var(v_t\ )$) because they adjust for the difference in means across levels of facet-a that drive the increase in bias in variance estimates. Further, models that adjust for facet-a will provide more estimates of the variance of teacher quality that contain less sampling variation (i.e. $var[\hat{var}(v_t^a)] < var[\hat{var}(v_t\ )]$), which is possible because the adjusted model eliminates the sampling variation in estimates caused by mean differences across the levels of facet-a. For example, a teacher will have the same estimated score from the adjusted model (i.e. $\hat{v}_t^a$) whether they were observed once, twice, or three times on facet-a_1 while the estimate of teacher quality from the unadjusted model will vary based on how

[5] Note that I am assuming here that within-teacher hidden facets are affecting within-teacher differences in teaching quality. If there is a correlation between a teacher's average teaching quality and the prevalence of a specific facet, the within-teacher facets can affect between-teacher differences in teaching quality. When this occurs, the within-teacher facets can lead to the same effects as between-teacher hidden facets because they take on a component that acts between teachers (which may or may not prevent generalizing across the facet, depending on the nature of the facet and the nature of between-teacher differences of the facet).

often they are observed on facet-a_1. Since few definitions of teacher quality are likely to treat sampling variation as an aspect of teacher quality, the adjusted model should be preferable in most cases, assuming of course, the shift in the definition of teacher quality resulting from using the adjusted model is acceptable. Importantly, the difference here stems from the fact that it is easier to accurately measure teacher quality within levels of the hidden facet than without regard to those levels. This reflects the same gain in efficiency that comes from stratified sampling as opposed to simple random sampling. Thus, adjusting models for within-teacher facets should reduce both the bias **and** variance in the estimate of the variance of teacher quality ($\widehat{var}(v_t)$). This should lead to better reliability of estimates of teacher quality when using the adjusted model.

II.2.3. Between-Teacher Hidden Facets Between-teacher hidden facets occur when teachers work in different contexts and the average of the observed teaching quality varies across context. When teachers are only observed on a single level of facet-a, little evidential basis exists from which to generalize scores across the facet or to support comparisons of scores between teachers across the facet. In general, two assumptions are commonly used to support extrapolation. The first, called teacher sorting, assumes that any observed differences in teacher quality across the between-teacher hidden facet are the result of true differences in teacher quality (i.e. $cor(v_t, student\ characteristics) \neq 0$). If the effect is completely due to teacher sorting, teacher quality should be comparable across levels of facet-a without the need for any adjustment (i.e. $\hat{v}_t = v_t$). Further, the adjusted model will provide incorrect estimates of teacher quality because the adjustment equates teacher quality across the levels of facet-a, whereas, by assumption, there are differences in teacher quality across the levels of facet-a (i.e. $\hat{v}_t^a \neq v_t$).

The second assumption assumes that any observed differences in teacher quality across the facet are the result of some characteristic of the facet (i.e. co-construction;

$cor\left(X_{ir(o:d:s:t)}, student\ characteristics|\upsilon_t\right) \neq 0$). The most common example of co-construction is that higher-achieving students may be easier to teach than lower achieving students, perhaps because they follow directions better or contribute more to the intellectual culture of classrooms. If the effect of facet-a is completely due to co-construction, only after adjusting for β_a will estimates of teacher quality be accurate (i.e. $\hat{\upsilon}_t^a = \upsilon_t,\ \ \hat{\upsilon}_t \neq \upsilon_t$). In this case, the difference across facets has nothing to do with a teacher's ability so the differences in observed teaching quality are, in a sense, artificial at least from the perspective of teacher quality.

It is generally difficult to empirically distinguish between teacher sorting and co-construction for facet effects, unless a study is designed explicitly to address this problem[6]. This is unfortunate because the implications of the two assumptions are contradictory. Thus, the assumption one makes in order to extrapolate determines whether the choice to adjust is correct.

II.2.4. Non-Constant Effects of Hidden Facet across Teachers In the previous section, I discussed how mean differences across levels of a facet can affect estimates of teacher quality. In many cases, however, differences in observed teacher quality across facets will be more complex. For example, if the facet is within-teachers (e.g. teaching writing or not), the effects of this within-teacher facet might vary across teachers. Suppose, for example, one group of teachers is particularly skilled in teaching writing, leading the difference in teaching quality for writing versus non-writing lessons to be much larger for that group. Another group of teachers might struggle with teaching writing, leading the difference in teaching quality for writing versus non-writing lessons to be very small for that group (i.e.

[6] It should be noted that I described only the two most common arguments for and against adjusting. However, many others (and combinations of the two posed) exist, which would suggest other methods of adjusting for observed differences across the facet. For example, I discuss the assumption that schools have a constant mean effect on scores across all teachers, which suggests centering scores within-schools as a solution in later chapters.

$\upsilon_t^{a_1} - \upsilon_t^{a_0} = \beta_{a,t}$ where $\upsilon_t^{a_1}$ is teacher quality for facet-a$_1$, $\upsilon_t^{a_0}$ is teacher quality for facet-a$_0$, and $\beta_{a,t}$ is the difference between the two and varies across teachers[7]). In this case, neither the unadjusted model nor the adjusted model would be fully appropriate (because the adjusted model estimates β_a and not $\beta_{a,t}$, which varies across can take different values across groups) and a researcher might want to consider more complex models. One approach would be to allow the β_a coefficient in the regression analysis to vary across the teacher random effect facet (i.e. υ_t). Alternatively, researchers could separately sample, from each teacher, days of instruction from each level of the within-teacher facet-a. This would allow separate teacher quality estimates to be estimated for each teacher at each level of facet-a. This sampling approach to correcting for the effect of a hidden facet on observed teaching quality creates facet-level specific teacher quality estimates for a given teacher (e.g. a separate teacher quality estimate for writing and non-writing). Arguably, estimating level-specific abilities for a facet can provide the basis for a richer exploration of the effects of a facet while making fewer assumptions about the data, though the approach requires much more data.

II.2.5. Hidden Facet Effects on the Variance of Teacher Quality To this point, I have focused on the effects of hidden facets on the average of the observed teaching quality (i.e. differences in means across levels of facet-a). In fact, facet-a can have effects beyond producing a mean difference in observed quality scores. Sampling each level of facet-a independently, as just discussed, would provide the most thorough exploration of this possibility. Full GTheory models, such as Equation (1), could be run for each level of facet-a. If this approach was used, the variance of any of the measurement facets, including the variance of "true" score (υ_t), could now vary across these independent models (i.e.

[7] Forgive the slight abuse of notation, which recasts $\beta_a$as a difference in teacher quality rather than observed teaching quality. I retain the notation of $\beta_a$to emphasize that I'm still referring to the same difference in quality across facet-a, though doing so at a different level of abstraction the two should be equivalent in the context in which its used.

$var(v_t|a_1) \neq var(v_t|a_0)$). For example, lectures may be a relatively simple form of instruction where teachers all have a fair amount of skill (i.e. $var(v_t|a_1)$ is small) while small group discussions may require far more from teachers and so better demonstrate a teachers' skill in teaching (i.e. $var(v_t|a_0)$ is large). This fact could be utilized to construct more reliable measures of teacher quality by measuring only facets with more teacher-level variation, but this, of course, would change the meaning of teacher quality and restrict generalization to the levels of facet-a observed. In order to explore these more complex effects of the hidden facets, it is necessary to observe teachers on multiple days within each level of facet-a. As I will describe later, that is not possible using UTQ data. Thus, in the empirical analyses presented in this thesis, I am restricted to examining only mean differences across levels of the hidden facets.

II.2.6. Role of Random Sampling in Analyzing Hidden Facets The discussion up to this point has assumed that, across teachers, the observed days are representative of the full universe of possible days. This is necessary to make any generalizations across the specific days observed and, as discussed above, is assumed by GTheory. There are two important parts of this assumption, however, both of which involve within-teacher hidden facets. First, the assumption is that sampling is ignorable, and preferably random. Sampling is ignorable if the likelihood of being observed on any level of a hidden facet is independent of teacher quality. Second, each possible level of the hidden facet is assumed to have a positive probability of being sampled. Importantly, if non-ignorable sampling occurs, any measurement errors discussed here may contribute to bias in scores while if some levels of a hidden facet have no chance of being observed, generalization cannot be done over that level of the hidden facet (though we could extrapolate).

Either of these assumptions can break down in practical settings. Observations within teacher evaluation systems, for example, face competing time demands from busy schedules.

Further, the goal of random sampling can conflict with formative assessment goals of evaluation systems. For example, a teacher may want to be observed only on writing lessons to get feedback on a type of instruction they find challenging. This could negatively bias scores for that teacher and make it impossible to generalize scores beyond writing lessons for that teacher. Further, assuming that teachers with higher teacher quality were more likely to engage in this practice, it would bias estimates of the effect of writing lessons by creating an association between teacher quality and the likelihood of being observed teaching writing (i.e. writing instruction will take on a between-teacher nature and be caused, at least partly, by teacher sorting). Thus, non-ignorability in the sampling of days is a significant challenge. This challenge, however, is likely minimal in research where efforts are made to keep sampling random and few incentives exist for teachers to manipulate sampling. The challenge, though, will likely be more important for accountability systems in practice, which must balance both the summative and formative goals of observations and complex schedules. This is an important fact because the effects of within-teacher facets that I detect in UTQ, which engaged in random sampling of days, may be a lower bound of the effects one might see in practice.

II.2.7. Hidden Facets and Construct Validity One last problem arising from hidden facets concerns the concurrent validity of estimated teacher quality from a given statistical model. Assume there is an alternate measure of teacher quality (τ_t), such as teacher value-added (VA) scores. The question discussed now is how statistical adjustments of teacher quality estimates from classroom observation data affect correlations of these teacher quality estimates to an alternative measure. This is an important question because the goal of measurement is teacher quality and, by demonstrating that teacher quality estimates, after making some adjustments for hidden facets, have greater concurrent validity, evidence is provided that these adjustments improve the measurement of the teacher quality construct.

I begin by discussing how adjusting for *within-teacher* hidden facets might affect this correlation. Based on the arguments to this point, adjusting for *within-teacher* hidden facets should increase the precision of teacher quality estimates. This increase in precision, in turn, should increase the correlation between teacher quality estimates and the alternate measure (i.e. $cor(\hat{v}_t^a, \tau_t) > cor(\hat{v}_t, \tau_t)$) and so give a “better” estimate of teacher quality. However, the alternative measure, the unadjusted model, and the adjusted model all have different ways of defining teacher quality. If the definition of teacher quality implicit in the adjusted model is more aligned to the alternative measure, this alignment could off-set the gain from increased precision. For example, suppose that content coverage by teachers has large effects on student achievement such that teachers who cover more writing will have better value-added scores (Polikoff and Porter, 2014). Under these conditions, the unadjusted estimate of teacher quality taken from classroom observation scores will capture the effects of any differences among teachers in content coverage (to the extent these differences are estimable from observed data) whereas the adjusted model removes these effects from estimates of teacher quality. Thus, the unadjusted model may be superior because it retains a piece of teacher quality that is important to value-added scores, though, as I discussed above, it is not clear that the frequency of teaching writing can be reliably estimated with observation scores. The overall point, then, is that adjustment for within-teacher facets can improve precision (and therefore should improve concurrent validity). However, unadjusted models could be better aligned to the alternate measures and this "alignment" effect could be larger than gains from precision. The net effect of these two forces (precision vs. alignment) is therefore hard to predict.

Adjusting for *between-teacher* facets, by contrast, mainly affects the bias with which teacher quality is estimated and this bias could either increase or decrease the correlation between the teacher quality estimate and the alternate measure. The correlation should increase if co-construction effects are at work but decrease if teacher sorting is at work. For example, if more advantaged students are easier to teach (i.e. co-construction), but are generally not taught by better teachers (i.e. no sorting), we might observe higher teaching quality in classrooms with more advantaged students. In this case, the adjusted model will have a higher correlation between estimated teacher quality and the alternative measure because only the adjusted model will accurately reflect that teacher quality is unrelated to the percentage of advantaged students in a classroom (unlike unadjusted measured teaching quality). On the contrary, if better teachers choose to teach more advantaged students (i.e. teacher sorting), we might observe the same higher teaching quality in classrooms with more advantaged students. However, in this case, the unadjusted model will have a higher correlation between estimated teacher quality and the alternative measure because only the unadjusted model will accurately reflect the true differences in teacher quality across classrooms with different percentages of advantaged students. Thus, by using a concurrent measure of teacher quality, it may be possible to test for an increase in precision when controlling for within-teacher hidden facets and to explore the role of teacher sorting and co-construction in explaining between-teacher hidden facets. However, concerns about alignment make this a difficult proposition.

Finally, the relationship of observed teaching quality and true quality may differ across different hidden facets (i.e. $cor(\mathrm{X}_{\{\mathrm{ir(o:d:s:t)}\}}, \tau_t|a_1) \neq cor(\mathrm{X}_{\{\mathrm{ir(o:d:s:t)}\}}, \tau_t|a_0)$). For example, if true quality in lectures is driven by the organization of the material and true quality in small group discussions is driven by instructional interactions, CLASS should be more valid for small group discussions because it measures interactions better than the organization of content. Thus, we might also be interested in the validity of estimated teacher quality across levels of facet-a. Of course, testing the validity of observation score estimates in this way presupposes that we actually have an alternate measure of teacher quality[8]. One challenge of finding such a measure, as I detail later, is ensuring that the measurement error in the alternative measure must be uncorrelated to measurement error in the estimated teacher quality from observation scores. This is a challenge because school context and students may lead to correlated errors across both measures. A further challenge is alignment of different definitions of teacher quality, as described above. Overall, then, there are many challenges to addressing the validity of estimates of teacher quality that will need careful qualitative and theoretical exploration. Additionally, experimental methods will likely be necessary to truly examine the validity of scores.

II.2.8. Summary At this point, let me review the implications of the discussion so far. I have argued that hidden facets exist and can affect observed teaching quality. Differences in observed teaching quality across levels of a hidden facet may be due to instrument bias or true differences in teaching quality, which in turn may be the result of within-teacher or between-teacher facets. Instrument biases will lead to biased estimates of teacher quality. Within-teacher facet effects on observed teaching quality will increase the sampling error of observed teaching quality, inflating estimates of the variance of teacher quality, but should not lead to bias *if* sampling is ignorable. Between-teacher facet effects on observed teaching quality may or may not bias estimates of teacher quality, depending on whether teacher sorting or co-construction are the source of these effects. It is possible to adjust models for mean differences in teaching quality across levels of the hidden facets, but this shifts the meaning of the teacher quality estimates. Further, differences in average scores

[8] Unfortunately, there is no good measure of teacher quality (τ_t) in UTQ. VA scores are too distal to detect anything but large effects (though I will test for effects with them). I cannot use the multiple observation instruments either because they correlate in similar ways to hidden facets (i.e. shared measurement error). Adjusting for hidden facets will remove this shared source of variation, necessarily decreasing the correlation between observation measures. A measure of teacher quality that does not share sources of error is necessary, instead.

across hidden facets are only the simplest of many possible effects of hidden facets. A better option, I have argued, might be to sample separately each level of the hidden facet, constructing a full GTheory model for each facet, which would allow a full exploration of the impact of the hidden facet. However, this is a data intensive approach that most data sets cannot support. The question of validity floats above this enterprise, but is very elusive. Without a good alternative measure of teacher quality, validity cannot be addressed, but identifying a good measure is complicated by the many sources of measurement error and complications stemming from shifting definitions of teacher quality across models. Nonetheless, the validity of estimates of teacher quality may either increase or decrease after adjusting for the effects of hidden facets and this change should provide information about why hidden facets are associated with observed teaching quality.

II.3.Three Classes of Hidden Facets

As we have just seen, hidden facets pose a challenge to measuring teacher quality and making accurate comparisons of teacher ability across contexts. In this section, I describe three classes of hidden facets. These classes are differentiated by how and why they affect observed teaching quality. I then use this distinction to further discuss when one may wish to adjust scores for the effects of these facets. The first class of hidden facets to be discussed includes facets of measurement stemming from the observational system in use. In what follows, I call these System Design facets. These are facets of measurement that are introduced by the necessity of selecting specific days, times, and raters to score teachers as part of an observation protocol and almost certainly contribute to measurement error. The second class of facets includes characteristics of Curriculum and Instruction that affect measured teaching quality. These Curriculum and Instruction facets can appear both within-teachers and between-teachers. Adjusting for these effects most directly changes the meaning of teacher quality estimates, shifting the definition of teacher quality to be a teachers' ability to teach *within* each level of the facet rather than generalizing teacher ability across all levels of the facet. The third class of facets discussed here arises from the organization of schooling. This class includes, for example, differences in the percentage of poor or linguistically disadvantaged students in a teacher's class. These contextual differences produce largely between-teacher effects on teacher quality estimates, and sometimes between-school effects, and when this occurs, an analyst needs to extrapolate in order to compare teachers across these facets.

II.3.1. Facets of System Design (SD) The first class of facets comes from the design of observation systems. This class captures differences in observed teaching quality arising from when observations are made, which classes are observed, and who is doing the observation and other facets of measurement associated with the observation system design. For example, systematic variation of teaching quality across the school year may occur because of structural features of classrooms (e.g. an initial honeymoon period of good behavior) or from structural features of the school environment (e.g. a focus on standardized testing in the early spring). Since observation systems organize when observations occur, the system determines whether teachers are observed across a range of time periods across the school year or during only a few time periods across the school year. This affects whether it is advisable to generalize teacher ability measures across the full school year and how the time of year facet affects estimates of teacher quality. When observational protocols are well-designed and implemented, these facets should mostly act within-teachers. I focus on this case, though it should be noted that when observational systems are not well-implemented, which can happen in practice, these facets may act between teachers.

Variation in observed teaching quality due to System Design facets occurs within-teachers, at least when sampling plans are well designed. Thus, these facets affect observation scores independently of teacher quality[9]. Based on the above discussion, this means these facets contribute to measurement error and may be sources of instrument bias when generalizing across the facets, but no extrapolation should be necessary since one could expect good sampling to produce observations across all or most levels of the facet. Thus, adjusting for the System Design facets in the measurement process should reduce sampling error, leading to more accurate score estimates. There are two main ways of adjusting for System Design facets, both of which were described above. The first is to add fixed effects of the facet to statistical models (e.g. β_a as above), controlling for differences in average scores across levels of the facet. The second, preferred method, is to stratify sampling such that each teacher is observed across the various levels of a facet. In fact, this already occurs for some, but not all, SD facets in most research studies, including UTQ. For example, observations are usually spaced across semesters, controlling for time of year effects. As discussed above, using sampling to adjust for facet effects is preferred because it does not require the assumption that the average effect on observed teaching quality is the only effect of the facet nor complex statistical adjustments. Importantly, the size of the effects on observed teaching quality that I detect will apply most directly to research projects that sample as carefully as the UTQ project did. The effects of these facets in evaluation systems, which face greater constraints on the sampling of observation days (including the problem of non-ignorable sampling), likely will be much larger.

Raters are the System Design facet that has received the most scrutiny in the measurement literature, and raters are included as a planned facet of measurement in most GTheory analyses. Three separate challenges related to rater error arise in observation systems. The first is the problem of rater leniency (i.e. rater severity or norming), which

[9] Note, however, that teacher quality might *interact* with some of these facets. For example, some teachers may be better at getting the school year off to a quick start or maintaining even quality across the school year.

involves getting raters to agree on what constitutes performance at each scale point. The rater main facet and rater-by-item facet in a GTheory statistical model capture these problems. The rater main effect captures each rater's expected deviation from the average score and the rater-by-item interaction captures the possibility that this expected deviation might vary by the item being scored. A second challenge is rater uncertainty, which results from raters scoring inconsistently across occasions, teachers, or class sections. The third challenge is rater bias, where raters respond to some quality-irrelevant aspect of instruction, leading a rater to produce scores that are systematically higher or lower than true quality. The rater-by-teacher/ rater-by-section/ rater-by-day/ rater-by-occasion facets and all three-way facets involving raters capture a combination of rater uncertainty and rater bias, which are difficult to distinguish. The two-way facets just discussed capture raters disagreeing over the correct score for the teacher, section, day, or occasion while the three-way facets allow this disagreement to vary across items. Taking steps to account for rater error in a GTheory statistical model is important because rater effects are often large (Casabianca, Lockwood, & McCaffrey, 2015; Kane et al., 2012). Despite this, studies sometimes ignore rater error (e.g. J. L. Brown, Jones, LaRusso, & Aber, 2010; Burchinal, Vandergrift, Pianta, & Mashburn, 2010; Curby et al., 2009; Curby, Rudasill, Edwards, & Pérez-Edgar, 2011; Hamre et al., 2013; Kane, Taylor, Tyler, & Wooten, 2011) or account only for rater leniency effects (e.g. Cor, 2011; McCaffrey, Yuan, Savitsky, Lockwood, & Edelen, 2014). By contrast, Equation (1) includes 10 different types of rater error, allowing for a much wider exploration of rater errors[10].

[10] All models are unable to detect rater errors that are shared across all raters in a sample. These arise when the rater group as a whole deviates in the same way from the "true" score that should be awarded--perhaps because of inadequate norming at the training stage. Such effects may be large, especially when there are few raters and raters work closely together. Moreover, such errors can arise when raters using the same instrument are trained by different trainers for different studies. Estimates based on calibration data collected by UTQ researchers (which estimate differences between group means in scoring and an expert-provided "true" scores) suggest that group effects may account for up to 50% of the rater error in the UTQ data set.

II.3.2. Facets of Curriculum and Instruction (CI) The second class of facets comes from the variety of instructional goals (e.g. introducing content, reviewing), the variety of content topics (e.g. reading, writing), and a variety of other factors (e.g. rigor of tasks, instructional grouping) that occur in a teacher's classroom over time. These are “hidden” facets of measurement when they are not included as a facet in the statistical model (i.e. they are "hidden" from the model), but when variation in these facets nevertheless affects observed teaching quality scores on an occasion of measurement. Variation in levels of these Curriculum and Instruction facets occurs within-teachers, and all teachers generally will engage in instruction across all levels of these facets, though teachers likely vary in the amount of time they spend on each level of each facet, introducing a between-teacher component to these facets.

Additionally, this class of facets likely gives rise to the most instrument bias, as the content taught and instructional approaches may change the relationship between observed quality and true quality. This class of facets, then, represents a more complex challenge than the System Design facets and adjusting for the effect of facets is likely to be controversial because the frequency with which teachers engage in instruction at different levels of these facets is often considered an aspect of teaching quality (e.g. Polikoff & Porter, 2014). For example, if the effects of content domain on observed scores are statistically controlled for in a measurement model, comparisons across teachers will reflect teacher skill within each content domain, not the average provided teaching quality. This represents an important shift in the meaning of teacher quality. As I have argued, this shift removes the threat of biases stemming from instrument bias and improves measurement precision at the expense of sacrificing an aspect of teacher quality (prevalence rates in instructional activities). Thus, even if one believes the prevalence with which teachers engage in specific types of instruction is an important aspect of teacher quality, it may be beneficial to adjust for CI facets because the gains in precision and reduction in instrument bias may outweigh any biases introduced by the adjustment. Further, as I have argued, observation instruments do not measure instruction frequently enough to estimate the prevalence with which teachers engage in specific types of instruction, which is necessary to estimate how the prevalence of instructional practices (which is ignored when adjusting for CI facets) affects teacher quality. That said, adjusting for curricular and instructional facets goes against the typical ways of framing and understanding teacher quality so it is likely to be controversial.

My strategy in this dissertation therefore is to estimate models with and without adjustments for facets of Curriculum and Instruction. The differences across models show the impact of shifting the meaning of teacher quality (at least for when sampling is nearly random), and demonstrate the gains or losses from adjustment. If there is no meaningful difference in parameter estimates across models (i.e. $\hat{\upsilon}_t \approx \hat{\upsilon}_t^a$), the problem of adjustment remains academic with little practical importance.

II.3.3. Facets of School Organization The third class of facets provides a different type of challenge because these facets always affect between-teacher differences in observed teaching quality[11]. This means that extrapolation is necessary to interpret teacher scores as applying across these facets (since teachers will almost always be observed teaching in only one level of each facet). The effect of School Organization facets on observed teaching quality can be within-schools (e.g. tracking between teachers) or between-schools (e.g. residential sorting of students, school culture). This set of facets likely includes both co-construction effects, where the facet enables higher observed teaching quality, and teacher sorting effects, where teachers choose where they work. Thus, as discussed above,

[11] Note again that I've theoretically sectioned off the within-teacher context effects (e.g. variation in students across sections, teachers teaching multiple subjects or grades, and/or teachers teaching within different programs within the school) because they are better thought of as problems of system design--when observation systems choose to observe teachers.

extrapolation arguments with contradictory implications are feasible. As an example, it is not clear whether statistical models that adjust for School Organization facets will provide better or worse estimates of teacher quality. Moreover, the preferred sampling approaches described earlier are not possible because teachers typically cannot be observed across a range of levels of these facets. Making adjustments risks over-correcting for true differences across teachers while making no adjustments risks penalizing teachers who teach disadvantaged students. Despite these challenges, calls are already being made to adjust estimates of teacher quality for the effects of this class of facets (e.g. Whitehurst et al., 2014). Additionally, differential validity and instrument biases may also play a role here. For example, there is evidence that different types of students benefit from different types of instruction (Connor et al., 2009b), implying observation instruments could give higher scores to instruction that only promotes learning for certain types of students.

Missing from the current conversation about this third class of facets is the difficulty of making clear teacher quality comparisons across schools. Schools provide an environment and culture that supports or constrains teachers. This complicates the comparison of teacher quality across schools by making it difficult to distinguish between teacher and school effects. Thus, it is especially difficult to justify extrapolation arguments that support comparing estimates of teacher quality across schools. One solution, discussed by some value-added theorists (Raudenbush, 2013; Reardon & Raudenbush, 2009), is to assume teachers are comparable to other teachers only within their own school, or possibly very similar schools. This reduces the extent to which extrapolation arguments must bridge wide differences across contexts, simplifying comparisons across teachers. However, it prevents comparisons of teachers' who work in different contexts, forcing a definition of teacher quality that is isolated to the teachers’ current school and possibly very similar other schools.

II.4.Summary and Research Questions

In summary, this dissertation addresses two broad problems related to the use of classroom observation instruments to measure teacher quality. The first problem involves generalizing across the many facets of measurement involved in the typical classroom observation study, including both planned facets and hidden facets, in order to get an estimate of teacher quality. The second problem arises from the need to extrapolate across contexts where no data exists to generalize. The GTheory approach discussed in this dissertation provides a framework for exploring these twin problems.

The statistical model I develop explicitly includes a true score and multiple error facets and is used here to understand the contribution of each to observed teaching quality. The relative contribution to variance in observed scores of these different facets in a GTheory model has important implications for how reliably observation instruments are measuring teacher quality. It also has implications for problems related to the effects of hidden facets. For example, if there is no within-teacher, between-day variation in observed teaching quality (i.e. the day facet effect is zero), then measurement conditions that vary within-teachers between-days cannot affect observed teaching quality and so they are not facets of measurement. Given the complexity of the GTheory model I shown in Equation (1), it is also important to consider the precision with which variance components are estimated.

Of course, the sources of planned variation in a measurement system are not the only factors that can affect observed teaching quality. GTheory models can also attempt to incorporate the effect of "hidden" facets of measurement (at least if these are identified). As this dissertation shows, this can be done in a number of ways. The most straightforward way is to incorporate parameters that equate observed teaching quality across levels of the hidden facets. I have argued that this process can be used to identify areas of instrument bias, to examine the effects of within-teacher facets on the precision of teacher score estimates, and to understand whether between-teacher facets bias estimates of teacher quality. Determining if any of these effects are happening is important to understanding the impact of hidden facets on observed teaching quality. The decision of whether to adjust for the impact of hidden facets is more complex, requiring the balancing of shifting meanings of the estimated teacher quality score and the problems caused by the hidden facets in unadjusted models.

A last issue discussed in this dissertation concerns the validity of estimates of teacher ability. I have argued that the precision of estimates of teacher ability should increase after adjusting for within-teacher facets, especially facets related to System Design and Curriculum and Instruction. This, in turn, should increase the concurrent validity of estimates that adjust for the facet effects (i.e. by improving the correlation of the ability estimate to an alternative measure of teaching quality). But I also argued that any gains to validity arising from adjustment may be outweighed by shifts in alignment between the adjusted measure and the alternative measure. Additionally, the concurrent validity of observed teaching quality may vary across levels of hidden facets. That is, observation instruments may more accurately measure true teaching quality at specific levels of specific facets. This, in turn, would lead estimates of teacher quality to be more valid (i.e. more highly related to a concurrent measure of teacher quality) when teachers are observed on some levels of a facet than when they are observed on other levels of the facet (e.g. teachers observed teaching writing may have more valid scores than those observed teaching reading).

The considerations just discussed lead to the following research questions to be discussed in this dissertation:

1. Using UTQ data from several classroom observation instruments, what percentage of variance in observed teaching quality scores is due to a true score component (υ_t in equation 1) and what percentage is due to error components?

    a. To what extent do these percentages of variance differ across the classroom observation instruments used in the UTQ data?
    b. How precise are the estimates of these variance components in UTQ data?

2. When hidden facets are analyzed in a GTheory framework:
    a. Is there any evidence that the observation instruments used in the UTQ study may be biased for some levels of identified hidden facets?
    b. Do the hidden facets affect observed teaching quality within-teachers or between-teachers? Do hidden facets affect between-school differences in observed teaching quality?
    c. How much does adjusting for the effect of hidden facets on observed scores change estimates of teacher quality and estimates of the reliability of teacher quality?
3. Does adjusting for the effect of hidden facets on observed teacher quality in the UTQ data improve the relationship between teacher quality estimates and teachers' value-added scores?
    a. Does the concurrent validity of observation scores vary systematically across teachers based on the level of the hidden facets over which they were observed?

Chapter III. Review

There has been a wide range of research on classroom observation instruments recently. The nature of this research has varied quite widely across specific instruments discussed in this thesis. Research on CLASS has largely involved evaluating interventions designed to change the climate of classrooms, with CLASS serving as the proximal measure of classroom environment. Research on FFT has centered on evaluations of district teacher evaluation systems. Research on PLATO has centered on how to measure and understand subject matter teaching quality. The combination of these strands has increased our understanding of observation instruments as measures of teaching quality, but leaves this knowledge unorganized and instrument dependent. We do not know how well results of research on one instrument might generalize across other observation instruments or across new samples of schools and districts. One challenge is the wide range of instruments. For example, the Early Childhood Environment Rating Scale-Revised (ECERS-R) focuses on the pre-K physical environment (Cassidy, Hestenes, Hegde, Hestenes, & Mims, 2005) and the TEX-IN3 captures literacy environment (Hoffman, et al., 2004). I restrict my review here largely to high-inference, behaviorally-anchored observation instruments that directly score quality of instruction. I do not review research on instruments that do not directly measure instructional quality, low-inference or behaviorist-oriented instruments, time-sampling instruments, instruments focused on specific discrete behaviors, instruments used only in pre-school, or older process-product instruments that are not used anymore[12]. This restricts my focus to instruments similar to those that are adopted by newer teacher evaluation systems.

[12] Most of these older instruments either fall into the other excluded categories also or have very few published studies that I was able to locate that are relevant to the discussion here.

However, most research to date on instruments like this has focused on only four instruments: CLASS, FFT, PLATO, and Mathematical Quality of Instruction (MQI; Hill et al., 2012b). Thus, by default, research using these four instruments forms the bulk of this review, though I will discuss other instruments where possible.

It is important to consider how representative these four focal instruments are to the broader group of classroom observation instruments focused on measuring teacher quality. While the four focal instruments represent a wide range of ways of characterizing teaching quality, they by no means represent the full diversity of instruments under consideration for use in teacher evaluations (nor a random subset therein). For example, the Marzano Art and Science of Teaching Framework (Marzano) has raters score only selected portions of the instrument and Thoughtful Classrooms (TC) has items that are only scored when they fit the lesson (Rowan et al., 2013). These unique features could have important impacts on how these observation instruments function, especially when considering rater error. Additionally, the focal instruments were designed to capture a broad range of activities and lessons while some instruments capture only specific types of instruction or one small aspect of the classroom environment. For example, the Instructional Quality Assessment (IQA) scores only discussions (Matsumura et al., 2006). This may lead the IQA to capture discussions better than any other instrument, but this comes with the downside of a very narrow scope. Overall, then, the instruments that I discuss here are a non-representative sample of the observation instruments in use in schools and research today. They are the focus simply because they have been featured prominently in research, due in part because of their broad applicability, which simplifies their use, but also because of who designed them, why they were designed, and where they have been adopted in practice. While the problems that I discuss should be relevant to all observation instruments, research is needed to understand how much specific results generalize across instruments.

Given these caveats, I turn now to reviewing prior research on observation instruments that provide high-inference scores of teaching quality. The flow of this chapter follows the previous chapter. I start by reviewing the research using Generalizability Theory, which provides a broad sense of the functioning of observation instruments. I then turn to consider the situated nature of teaching, reviewing aspects of the measurement context that have been shown to affect observed teaching quality. This discussion is organized by the three classes of facets introduced in the previous chapter. Where research exists, I look to see how each facet might contribute to instrument bias or the reliability and validity of estimates of teacher quality. Last, I discuss what is known about the validity of observation scores as measures of teacher quality.

III.1. Generalizability Theory with Observation Instruments

Generalizability Theory (GTheory) has been the main tool for understanding the measurement properties of classroom observation instruments. Nine recent studies have conducted GTheory analyses on observation instruments that fit my criteria of being high-inference, behaviorally-anchored, direct measures of teaching quality (Bell et al., 2012; Hill et al., 2012b; Ho & Kane, 2013; Kane et al., 2012, 2011; Mashburn et al., 2013; Newton, 2010; Praetorius, Lenske, & Helmke, 2012; Praetorius, Pauli, Reusser, Rakoczy, & Klieme, 2014).[13] These studies covered 6 instruments, used 7 different statistical models, and organized scoring in a number of different ways, making it difficult to generalize conclusions across instruments or studies. One problem lies in the fact that none of the studies under review used a full GTheory model (of the sort presented in Equation (1)). This is one reason comparisons across studies are difficult (Shavelson, Webb, & Burstein, 1986). The most

[13] Note that I do not include process-product research from the 1980s-1990s because these studies were generally either focused on time-sampled instruments, low-inference instruments, instruments that did not directly measure teaching quality, or instruments that appear to not be in use anymore. Further, I wanted to focus on modern instruments, which tend to be more constructivist and/or socio-cultural than older behaviorist-leaning instruments.

common deviation from a full GTheory model involved averaging across items before analysis, which I call item-average models (Bell et al., 2012; Hill et al., 2012b; Ho & Kane, 2013; Kane et al., 2012, 2011; Mashburn et al., 2013; Newton, 2010). While there is nothing statistically incorrect with this, it complicates comparisons across studies and hides the effect of items. Item-average models have less overall variance because scores are averaged across items (thus reducing total variance) and have fewer modeled error facets. This results in a greater percentage of variance in item-averaged observation scores to appear to come from the remaining main facets (e.g. the teacher true-score variance and the variance of the rater error facet appear larger in item-averaged models). Thus, when comparing the amount of total variance due to the teacher true-score across studies that use different models, it was difficult to determine how much of the difference in results between studies was due to different samples and how much was due to the different models. The same applies to the other facets.

In addition to aggregating over items before conducting a GTheory analysis, two of the studies averaged across occasions of measurement within days before running models (Hill et al., 2012b; Kane et al., 2012). Aggregating over occasions within days before analysis inflates day variance while reducing the total variance, inflating the percentage of variance due to teachers, days, and raters. The MET analyses went one step further and used only main facets in their model (Kane et al., 2012). This ignores the rater-by-teacher and rater-by-section facets. The full impact of ignoring these facets is unclear, though it will inflate the residual variance estimates. Overall, the choice of which facets to model and whether to aggregate scores before analysis will change the variance associated with each modeled facet, making it difficult to know whether differences across studies are driven by model or sample.

Another important difference across studies that might affect the variance components is the level at which raters are assigned (i.e. assigned to score occasions, days, sections, teachers). Most studies assign raters to days so that raters score all occasions within a given day. One study assigned raters to occasions, such that each rater scores only one occasion on a given day (Mashburn et al., 2013). This latter study procedure resulted in higher estimates for occasion and rater-by-occasion variance and lower estimates for day and rater-by-day variance. This shift from day to occasion variance likely will always occur under this different rater assignment procedure. When raters are assigned to occasions, the rater-by-occasion effect captures two raters disagreeing on an occasion's score and the occasion effect captures the occasion's deviation from a day score composed of many different raters' scores (i.e. it includes some rater disagreement). When raters are assigned to days, the rater-by-occasion effect captures the raters' different perception of how the occasion deviated from their own view of the day score while the occasion effect captures average of raters' views of how the occasion deviated from their own estimate of the day mean (i.e. it includes no rater disagreement). Because raters disagree with each other, the net effect of this difference should be to increase the variance of planned error facets at the level at which raters are assigned. The level at which raters are assigned will capture the majority of rater effects stemming from stable rater disagreements. This is also true when raters are assigned to teachers, which appears to increase estimates of the rater-by-teacher error facet (Ho & Kane, 2013). Fully-crossing raters should help remove these effects, making the interpretation of error facets more clear.

Despite the above caveats about inconsistencies across studies, there is a great deal of consistency in study results. Teacher variance in observed scores was generally near 25-30% of the total variance and slightly higher when data came from practice. The higher teacher variance in practice appears to stem from principals using information not contained in the observation in their scoring (Whitehurst et al., 2014). Day variance estimates were more variable across studies ranging from approximately 10-20% of the total and noticeably higher in analyses of the MET data compared to other studies. Estimated rater effects showed the most variability across studies, which is to be expected given the different raters and training approaches across studies. Last, residual variance was always one of the largest variance components of observation scores. Overall, however, it is unfortunate that there is not more consistency in the statistical models used, nor any sense of the uncertainty in individual estimates, across studies. A greater level of consistency would allow more precise analysis and comparison of results than is possible presently.

III.2. Facets of Measurement

The GTheory models used in previous studies aimed to examine how teaching quality varies across measurement error facets like sections, days, occasions, raters, and items, which are the planned sources of generalization for classroom observations. But teaching is a situated task, and teaching quality will therefore vary systematically over many other aspects of lessons and classes as well (Gitomer & Bell, 2013). Understanding this variation is key both to understanding the generalizability of estimated scores and knowing the extent to which scores can be extrapolated across contexts. I turn now to a review of what we know about potential hidden facets. I divide this section into three parts based on the three classes of facets that I introduced in the last chapter. Very little past work has incorporated an exploration of these facets into a measurement framework, so the work that I review generally demonstrates simple mean differences in observed teaching quality across levels of a facet. Where possible, though, I will discuss whether research suggests that the hidden facets may be a source of instrument bias or affect the bias and reliability of estimated teacher scores.

III.2.1. System Design Facets There are two main elements of the design of an observational system that are likely to influence observed teaching quality. The first is the decision of when observations occur. The second is the choice of raters doing the observation. I also discuss the role of specific items in this section. Item effects are an important part of understanding observation instruments, especially as some districts are adapting instruments by changing, adding, or removing items (e.g. Chaplin, Gill, Thompkins, & Miller, 2014). An observational system must make decisions along each of these three areas to structure observations. These choices will impact estimates of observed teaching quality and affect how teaching quality relates to teacher quality.

III.2.1.1. When to Observe In this section, I discuss what is known about how the timing of observations affects teaching quality. Observed teaching quality fluctuates both randomly and systematically throughout the school year, the school day, and both within and across lesson periods. When teachers are observed at different points in time, the facets that effect observed teaching quality act differently across teachers. This complicates estimating teaching quality because part of the variation across teachers is due to the factors related to when teachers were observed. I discuss in this section the known factors that lead to this systematic variation across time, including occasions within days, time of year, time of day, and sections.

Occasions. Lesson periods are usually the focus of observations due to the natural division of a school day into lesson periods. Many instruments, though, do not assign scores to lesson periods, but instead break lesson periods down into shorter occasions, usually using equal-length occasions (e.g. CLASS scores 15 minute occasions). The division into occasions has received little explicit discussion or focused empirical study. Occasions must be long enough to provide evidence to score each item, but short enough to reduce the cognitive burden of scoring. The longer the time period being scored, the more raters must internally aggregate over many pieces of, possibly conflicting, evidence (Hill et al., 2012a). This is a complex cognitive challenge for raters, which may contribute to the high amount of rater error found in observation instruments. Scoring shorter segments should reduce the cognitive burden on raters, but may lead to scoring based on heuristic approaches if little evidence for an item is observed (e.g. assigning a score that matches the other items rather than scoring items individually; c.f. Bell et al., 2014). Further, it shifts the internal process of aggregating scores over time to an external, testable process of averaging scores across discrete occasions. This allows a more formal approach to combining scores from parts of a lesson into a whole, but may lose contextual considerations raters can employ internally when aggregating scores. Similarly, when scoring occasions live, the rater is recording scores for up to one-third of the lesson, which may lead rare events to be missed.

Given these considerations on occasions, I turn now to review how research has explored the role of occasions. Work on the Measures of Effective Teaching Project (MET; Kane & Cantrell, 2010) focused on the question of how many occasions are necessary to accurately estimate the average day score (Joe, McClellan, & Holtzman, 2014). Two fifteen minute occasions sufficed to get scores that correlate with the total score above 0.9, leading MET to score only the first 30 minutes of each day, though they acknowledged rare events may be missed by such a procedure[14]. Using FFT, another MET study found that 20% of domain 2 and 40% of domain 3 scores on this instrument changed when scores were given on the first 15 minutes rather than the whole day. This change was enough to shift teacher scores significantly up or down the distribution (Ho & Kane, 2013), suggesting one occasion is not sufficient. The other way of examining occasions has been to include occasions as a facet in GTheory analyses (Bell et al., 2012; Malmberg, Hagger, Burn, Mutton, & Colls, 2010; Mashburn et al., 2013), though none of these studies included a variable for the n^{th} occasion within a day (i.e. include a unique effect for the first/second/third occasion in a lesson period). Mashburn and colleagues (2013) found that occasion variance dwarfs day variance,

[14] The total score was the average across occasions 1-4 so this correlation is inflated by fact that the first two occasions composed half of the total score.

though their assignment of raters to occasions changes the meaning of the occasion facet (as described above). More commonly, occasion variance is found to be slightly less than day variance (Bell et al., 2012; Malmberg et al., 2010).

None of these studies get at the root of how occasions affect observation scores. To do so requires explicitly recognizing the ordering of occasions and the way they are created from a full lesson. There has been some recognition of occasions in this way. Minutes 15-30 in a lesson score higher on PLATO than the first fifteen minutes while time after the first 45 minutes is scored significantly lower (Cor, 2011). Other work found the first occasion of a lesson scores higher than other occasions (Cortina, Miller, McKenzie, & Epstein, 2015), though this may vary across items with instructional items increasing over the course of a lesson (Ho & Kane, 2013). Thus, there is inconsistency in the literature. This inconsistency may stem from way occasions are created (e.g. occasion 1 can start with the bell or when instruction begins).

The division of lessons into 15 minute occasions is arbitrary and leads to occasions with little coherence (Hill et al., 2012a; Staub, 2007). A proper examination of occasions would be served by creating meaningful occasions within lessons. Though this can be difficult, breaking lessons into occasions with a constant content focus and grouping structure has proved useful (Carlisle, Kelcey, Berebitsky, & Phelps, 2011; Stodolsky, 1984). Researchers from the Third International Mathematics and Science Study (TIMSS) argued instead for the use of lesson events, patterns of regular behaviors of consistent and pre-defined form and function that occur within cultures (Clarke et al., 2007). Adopting occasions with a meaningful structure introduces a new source of error: disagreement over the demarcation of occasions. On the other hand, making occasions coherent may reduce rater error by allowing raters to score a coherent piece of instruction rather than cognitively balancing multiple distinct pieces of instruction (Schutz & Moss, 2004). Additionally, it could better structure feedback for teachers and make more clear how teaching quality varies within lessons. At this point, however, no studies of relevant observation instruments have used coherent occasions to empirically test the impact. Thus, while the current literature suggests an important, albeit minor, impact of occasions on observed scores, the way occasions are studied may drive this finding, possibly reducing the apparent influence of occasions on observed scores. This is particularly relevant for this dissertation because it restricts how carefully I can study the effect of specific instructional practices on teaching quality.

Time of Year. The structure of the school year also leads to fluctuations in observed teaching quality. At the beginning of the school year, students and teachers are unfamiliar with each other. As they gain familiarity and establish instructional routines, interaction patterns may shift, leading to changes in observed teaching quality. Many studies have noted systematic variation in teaching quality across the school year. But this usually manifests as a linear decrease in scores over a semester (Pianta & Hamre, 2009) or year (Bell et al., 2012; Casabianca et al., 2013). This decrease varies across items, with classroom management items remaining more constant than other items (Bell et al., 2012; Casabianca et al., 2013). This decline may also be sample specific as beginning teachers may show gains in observed scores over the course of the school year (Kane et al., 2011; Malmberg et al., 2010). Few explanations of this decrease exist, though standardized testing may contribute, as teaching quality shows a marked decline just before testing (Plank & Condliffe, 2011). This research showing time trends has, for the most part, been conducted using CLASS and there is less evidence on how other observation instruments vary across the school year.

Time of Day. Scores appear to vary over the course of the day too, though evidence is limited. Most evidence suggests that teaching quality decreases over the course of the school day (Curby et al., 2011; Pianta & Hamre, 2009; Plank & Condliffe, 2011). However, this decline may be limited to specific dimensions as climate items appear to remain more constant (Plank & Condliffe, 2011). Examination of the effect of time of day has been conducted almost exclusively with CLASS and in lower elementary grades limiting its generalizability.

Sections. Teachers often teach multiple classes. In the elementary grades, they teach multiple subjects to the same students. In the upper grades, they teach multiple groups of students, possibly across different subjects and grades. This represents a possible source of variation in teaching quality (Bell et al., 2012). Observational systems must decide which sections to sample. This decision may have a large impact on conclusions about teacher quality. The difference between sections can result from differences in the students being taught or the subject being taught, which can occur either within-teachers or between-teachers, a distinction that is rarely taken up in the literature. Only when these effects are within-teachers do they belong to the set of System Design facets. Because these impacts are generally treated as between teacher effects, I discuss the role of students and subjects in the System Design facet section.

III.2.1.2. Raters Rater error has received more attention than any other source of error in observed scores. Further, an extensive literature exists about rater error in performance assessments more broadly. The two most consistent conclusions are that rater errors are much higher than desired (Bell et al., 2014; Cash, Hamre, Pianta, & Myers, 2012; Gitomer et al., 2014; Hill et al., 2012b), often exceeding accepted rules of thumb for rater errors (see Graham, Milanowski, & Miller, 2012), and an entire system of training, monitoring, and supporting raters is necessary to obtain accurate scores (Hill et al., 2012a; Joe, Tocci, Holtzman, & Williams, 2013). In fact, the high inference, global ratings of many classroom observation instruments have long been known to have high rater errors, even after extensive training (Hoyt & Kerns, 1999; Shavelson & Dempsey-Atwood, 1976).

The most thorough exploration of rater errors was conducted by Bell and colleagues (2014) using combined data from MET and UTQ. They found, both empirically and from rater self-report, that rater error varied by dimension with more dynamic dimensions containing more error while classroom management dimensions contained less error. Cognitive interviews revealed that all six raters under study displayed confusion in their *understanding* of at least one item, which went beyond trouble *scoring* the item from video data. This was despite their extensive training, calibration, and scoring experience. This confusion may stem from trouble reconciling discrepant beliefs about good instruction (Cash et al., 2012). The fact that raters misunderstood items is disconcerting as it suggests the assigned scores may not always reflect their intended meaning (Hill et al., 2012b), which may be even more true in practice because administrators are both less focused on accuracy and less experienced in using observation instruments than professional raters (Ferris, Munyon, Basik, & Buckley, 2008).

In fact, most results on rater error are overly optimistic because they examine error based on comparing scores given by two raters in order to estimate the amount of rater error. These are "internal errors" and can be contrasted with "external errors", where rater scores are compared to an externally created "true score" (Myford & Wolfe, 2009). This is an important distinction because the process of training and calibration may lead the entire group of raters to drift from the "true scores". Using the master scores from UTQ calibration data as a proxy for "true score", I found (in work in progress) that only about half of the rater error can be detected by comparing scores between raters. This can tentatively serve as an estimate of how much the current literature may under-estimate the actual amount of rater error in observation scores. However, the number of raters and how closely they work together will play a role in the prevalence of internal and external rater errors.

The most commonly examined rater error is rater leniency (i.e. rater main effect or υ_r from Equation (1)). Rater leniency arises when raters disagree in their understanding of how scale points correspond to actual performance. Estimates of rater leniency vary quite a bit across studies and instruments, ranging from 5% to 30% of the total variance (Bell et al., 2012; Hill et al., 2012b; Ho & Kane, 2013; Kane et al., 2012, 2011; Mashburn et al., 2013; Newton, 2010; Praetorius et al., 2012, 2014). While rater leniency is generally treated as constant, evidence shows that raters drift over time in how they use scales, causing leniency to vary over time (Casabianca et al., 2015; J. J. Cohen & Goldhaber, 2016; Harik et al., 2009). Leniency also varies across items (e.g. υ_{ir} from Equation (1)), which leads to covariances in rater error across items (Hoyt & Kerns, 1999; McCaffrey et al., 2014). These covariances make it difficult to explore correlations across items and the factor structure of items.

The most important determinants of rater leniency are the rater's role and goals (Golman & Bhatia, 2012). When administrators give ratings, scores are usually inflated and have a compressed range (Golman & Bhatia, 2012). Principals are especially sensitive to score thresholds that have consequences for teachers (Grissom & Loeb, 2016). This is because scores will affect the working climate, the principal's relationship with the teacher, and even whether the teacher takes up feedback provided (Bretz, Milkovich, & Read, 1992; Kraft & Gilmour, 2016). That is, principals are not solely focused on providing accurate scores, but have competing goals that influence how they score (Grissom & Loeb, 2016; Wang, Wong, & Kwong, 2010). Because of these challenges, districts sometimes put into place external observers with no connection to the teacher or goals beyond providing an accurate score (Dee & Wyckoff, 2015; Steinberg & Sartain, 2015). The scores derived from this procedure may be more highly correlated with VA scores, but are rarely as reliable (Ho & Kane, 2013; Whitehurst et al., 2014). The lower reliability occurs because principals are more familiar with their own teachers and so provide scores that are more stable (i.e. reliable) based on their knowledge of the teacher that extends beyond the lesson observed (Hoyt & Kerns, 1999). When this knowledge of teachers that extends beyond the specific lessons observed is not linked to teachers' VA scores, the validity of scores assigned by principals is lower than that of an external observer.

The rater leniency effects just discussed are only a small portion of total rater error, however. For example, MET found residual error, which is driven by rater inaccuracy and other rater errors, to be up to 10 times larger than leniency effects (Kane et al., 2012). This rater inaccuracy error stems from raters applying the observation instrument differently to specific videos. It also tends to be item specific with dynamically scored items and instruction-related items leading to higher rater error (Bell et al., 2014; Gitomer et al., 2014; Sartain, Stoelinga, & Brown, 2009). While rater inaccuracy is treated as purely random variation in scores caused by raters, little of the error is likely to be truly random (Murphy & Deshon, 2000). Rather, it stems from various sources, such as the way raters sample interactions from videos (Bell et al., 2012), the rater's current emotional state (Floman, Hagelskamp, Brackett, & Rivers, 2016), and even the previously watched video (Ho & Kane, 2013; Sumer & Knight, 1996). Rater biases against specific types of instruction or specific teacher and student characteristics may also play a role, though efforts to reduce bias are generally employed in rater training (Park, Holtzman, & Chen, 2014).

Rater error is an inevitable aspect of observation scores and likely to remain a major source of error. The research shows this consistently, despite extensive training and monitoring of raters. Rater errors in practice are likely to be higher, given the competing demands on and goals of administrators. However, it is important to note that the most damaging forms of rater error, biases against specific groups of teachers or students, have not been reported, though, admittedly, research designs and analyses able to capture these biases, if they exist, are uncommon.

III.2.1.3. Items The role of items has received relatively little attention because the focus of classroom observation instruments has largely been on identifying effective teachers using average scores. The main result in regard to items has been that average scores on items related to classroom management and culture are generally higher than average scores on items related to instruction (e.g. Kane et al., 2012). This shows that average scores on observation instruments will vary greatly based on how many items measure classroom management and culture compared to how many items measure instruction. In fact, the variance in classroom observation scores due to items dwarfs the variance from any other source (White, 2017). Some studies have explored items more carefully, especially early studies of the subject-specific instruments (Grossman et al., 2013; e.g. Hill et al., 2012a) and studies that present item-specific variance decompositions (e.g. Kane et al., 2012). Such studies have provided information about the validity of specific items. Explicit Strategy Instruction from PLATO, for example, shows the strongest relationship with VA scores compared with other PLATO items (Grossman et al., 2013). Studies also demonstrate that specific items show very different amounts of variation across raters and days, suggesting the need to explore item-specific models. The instructional triangle (D. K. Cohen & Ball, 1999) provides one way to understand this variation. Across days, teachers and students remain the same while content shifts so items connected to content should show greater variation across days, a consistent finding, though differences are often small and determining which items vary with content is fraught (Praetorius et al., 2014).

This research shows that the choice of what items are included on an observation instrument has an important effect. Further, items do not function in similar ways; some items provide more reliable estimates of teaching quality than others. Additionally, items scores vary in the degree to which they correlate with value-added scores. These differences across items, in part, are aligned with the distinction between items focusing on management or culture and those focusing on instruction. However, more work is needed to explore whether specific types of items function in unique ways across different instruments.

In summary, then, this section discussed the facets of system design. These facets are affected by decisions made in the design observational systems. There is a convincing body of evidence demonstrating that these facets affect observed teaching quality, though the extent to which these effects may be sample or instrument dependent is less well known. Well-designed observation systems space observations across time, hire well-trained raters who display few biases, and record data on teaching quality across a wide range of dimensions. In these well-designed systems, the effects of SD facets on observed teaching quality discussed here should produce minor sources of sampling error, inflating estimates of the variance of teacher quality. This will affect how reliably scores are estimated, but scores should not be biased by these facets.

III.2.2. Curriculum and Instruction Facets Features of curriculum and instruction can affect the observed teacher quality. Over the course of the year, teachers teach multiple content areas and use many instructional formats, such as lecture, recitation, and discussions. Understanding the extent to which these facets affect observed teaching quality and whether classroom observation instruments idiosyncratically respond to quality-irrelevant features of the enacted curriculum and instruction is important to building models that accurately estimate teacher quality (Brophy, 2006; J. J. Cohen & Goldhaber, 2016; Grossman et al., 2010; Kelcey & Carlisle, 2013). The most direct way to explore the potential effects of enacted curriculum and instruction on observed teaching quality would be to observe all teachers teaching the same lesson. Such analyses used to be more common in research on teaching and were used to increase the precision of measurement by restricting the variation in classroom tasks (e.g. Calkins, Borich, Pascone, Kugle, & Marston, 1977). This same approach provided useful information about how much specific tasks affected observed classroom quality scores. Current research, however, has been more focused on generalizing observed teaching quality scores to average provided instruction, and as a result, approaches that allow an exploration of how specific lesson plans affected observed teaching quality have fallen out of favor.

III.2.2.1. Content Domain Effects There is evidence that the content domain taught affects observed teaching quality (Grossman, Cohen, & Brown, 2014; Grossman et al., 2014). By content domain, I refer to large categories of content, such as reading, writing, grammar, fluency, or vocabulary in English and algebra or geometry in math. In a small study of English using the PLATO instrument, Grossman and colleagues (2013) found that lessons involving writing instruction received lower scores than lessons involving reading instruction, which the authors assert was due to the low direct instruction and the high frequency of seat-work practice in writing lessons. Using data from the larger MET study, though, Grossman, Cohen, and Brown (2014) found that lessons involving grammar and lessons that involved both reading and writing received lower scores on PLATO than did lessons involving only reading or involving only Writing. This second study, then, failed to fully replicate the original study (Grossman et al., 2014). One limitation of both of these studies is that they have only examined the PLATO observation instrument, which focuses on English specific instructional practices. It is not clear if more general observation instruments (e.g. FFT or CLASS) will show these same effects. It is also not clear how much the differences in findings across the two studies just discussed is due to sampling error stemming from the use of different samples or due to other factors such as the writing curriculum used by the schools.

Overall, however, the finding that content domain affects observed teaching quality is not surprising given that teachers engage in different instructional moves across content domains (Stodolsky, 1984). In fact, other research (on reading instruction in elementary grades) found that the content domain being taught during a reading lesson accounts for about 15% of the variation in teacher moves associated with delivering instruction and 67% of the variation in teacher moves associated with supporting students (Kelcey & Carlisle, 2013).

There is a limited amount of evidence for content domain effects on teaching quality in Math. Indeed, I was able to find only one study that explored content domain effects on observed teaching quality in math (Hill et al., 2012b). This study found no difference in MQI scores across lessons focused on Algebra versus Geometry. Thus, more study is needed to examine the potential effects of content domain on observed teaching quality, both across new samples and using a wider range of instruments to help clarify how facets related to Curriculum and Instruction affect observed teaching quality.

III.2.2.2. Structure of Instructional Interactions The way instructional interactions are structured during lessons may also impact observed teaching quality. For example, individual seat-work limits interactions, recitations limit students' interactions to responding directly to teacher questions, while group discussions allow for more free-flowing and complex interactions among students and teachers. The choice between instructional formats structures the sort of interactions likely to occur among students and teachers, which in turn could affect observed teaching quality. While the choice of how to structure interactions is often considered a part of teaching quality, within-teacher variation in this choice will be observed as a result of sampling of days, which can affect the precision of estimates of teacher quality. Thus, the effect of how instructional interactions are structured on observed teaching quality may be either within-teachers or between-teachers.

The only interaction structure that has received research attention is instructional grouping. Small group work promotes more student to student interactions while individual work promotes fewer interactions. There is a range of findings, almost solely for CLASS, with some studies finding that individual work is rated lower on average than small group and whole class instruction (Curby et al., 2011; Plank & Condliffe, 2011, 2013). But other studies find no effect of grouping (Rimm-Kaufman, Paro, Downer, & Pianta, 2005; Stuhlman & Pianta, 2009). The difference across studies may be explained by the purpose of the individual seat-work in a given lesson, as individual work geared towards standardized test preparation is of particularly low quality (Plank & Condliffe, 2011, 2013). Thus, additional research is needed to understand the role that variations in interaction structures plays in the measurement of teaching, with a particular need for research focused on understanding whether effects generalize across instruments, grades, specific content being studied, and a broader range of interaction structures.

III.2.2.3. Other Curriculum and Instruction Facets Other facets related to the curriculum might also play a role. While there is no empirical evidence for this, a number of areas have been pointed to as potentially important to examine. For example, the sequence of content and lessons has long been highlighted as an area for study (Gage & Needels, 1989; Garrison & Macmillian, 1984; Staub, 2007). A lesson's learning goals have also been suggested as relevant to observed instructional quality (Kelcey & Carlisle, 2013). Some instructional goals may not require cognitively demanding instructional practices, leaving items focused on cognitive demand, such as Analysis and Problem Solving in CLASS, to be less valid for these lessons (Grossman et al., 2014; Praetorius et al., 2014; Walkington & Marder, 2014).

Overall, then, there is limited evidence about how facets of Curriculum and Instruction affect teaching quality. Content domain seems to be important, but the consistency of findings across samples and instruments is unknown and few other facets have been actively explored. However, teaching a full curriculum over the course of a year necessitates a wide range of teaching practices. We know very little about how this variety might systematically impact observed teaching quality or the inference to teacher quality, but it seems unwise to assume a priori that these effects are trivial. As such, we should build up our understanding of how observed teaching quality varies across the full range of curriculum and instructional practices.

III.2.3. School Organization Facets Yet another set of influences on observed teaching quality arise from the ways schools are organized. In this section, I review the research on four such facets and how they might affect observed teaching quality. The first facet is related to student characteristics, which vary within and between schools due to tracking and residential sorting. The second and third facets I discuss are subject and grade, which arise from the division of schools into discrete classes and grades. Lastly, I look at the impact of schools and districts overall on observed teaching quality.

III.2.3.1. Student Characteristics Studies consistently find that observed teaching quality is related to students' prior achievement (Allen et al., 2013; J. J. Cohen & Goldhaber, 2016; Polikoff, 2015; Schacter & Thum, 2004; Steinberg & Garrett, 2016; Whitehurst et al., 2014). These effects appear to be stronger for English compared to math, in middle schools compared to elementary schools, and for dimensions of teaching quality that relate to climate and culture versus instructional practices (Gill, Shoji, Coen, & Place, 2016; Lazarev & Newman, 2015; Steinberg & Garrett, 2016). In fact, the relationship between student characteristics and student's prior achievement may disappear entirely in elementary schools (Lazarev & Newman, 2015; Steinberg & Garrett, 2016), a finding that needs further study. The relationship of observed quality and student demographics is also well established, though the effect is weaker than for prior achievement (Bell et al., 2015; Chaplin et al., 2014; J. J. Cohen & Goldhaber, 2016; Grossman et al., 2014; Walkington & Marder, 2014).

Interestingly, teacher scores varied little across different groups of students taught by the same teacher within a given year, as when teachers offer instruction to multiple class sections in the same school (Kane et al., 2012). This seems at odds with the strong effect of student characteristics on observed teaching quality scores[15]. One explanation for the finding of small class section effects on observed teaching quality is that effects due to student characteristics are only between-teacher effects, which appears to be true for evaluation systems, where student characteristics largely act between schools (Jiang & Sporte, 2016; Kane et al., 2011). However, in the MET data, at least, where section effects are small, the relationship of prior achievement and observed quality were estimated as a within-teacher effect (Steinberg & Garrett, 2016). Another explanation for the puzzling finding about class section effects is that while student composition effects are present, they have relatively little effect on scores as a whole, at least when differences in student characteristics are measured as differences between sections taught by the same teacher. This was true in the MET data, where correlations between teacher score estimates with and without adjustments for student demographics were above 0.9 (Kane, McCaffrey, Miller, & Staiger, 2013). Further, the cross-year instability of observed teaching quality scores was found to be mostly unrelated to changes in demographic characteristics of classrooms (Polikoff, 2015). Both of these findings, then, suggest that the relationship between student composition and observed

[15] The author's own analyses show that in MET and UTQ data, there is a significant amount of variation in student characteristics within-teachers across-sections. This reduces the likelihood that between section differences in student composition are too small to detect, ruling out the possibility that teacher sections are too similar to detect student composition effects.

teaching quality is not highly important to estimating teacher scores overall, at least for some samples. That said, there is evidence that teacher behavior changes as a result of student composition (e.g. Carlisle et al., 2011) and that ideal instruction varies across students of different ability levels (e.g. Connor, Morrison, & Petrella, 2004; Connor et al., 2009a). Because of this, the relationship of student characteristics to teaching quality should not be ignored. Further, the effect on observed teaching quality of adjusting scores for student composition effects appeared to be much larger in many studies other than studies using the MET data (Jiang & Sporte, 2016; Kane et al., 2011; Whitehurst et al., 2014). These considerations have led to calls for adjusting observation scores based on student characteristics (Steinberg & Garrett, 2016; Whitehurst et al., 2014). However, as discussed earlier, any adjustments in observed scores for student composition can result in increasing errors in observed scores, unless the mechanism responsible for non-random assignment of students across classes is correctly modeled (J. J. Cohen & Goldhaber, 2016).

Beyond the direct effect of student characteristics on observed teaching quality, there is some evidence that the effect of observed teaching quality on student learning varies across different groups of students. A number of studies have found that lower ability students (Cadima, Leal, & Burchinal, 2010), poor students (Carlisle et al., 2011), and minority students (J. J. Cohen & Grossman, 2016) benefit more from high quality instruction than do their better off peers. Other work, almost entirely from the CLASS instrument, shows that observed teaching quality acts more to buffer students at risk of negative outcomes than to explain positive outcomes (Cadima et al., 2010; Curby, Rimm-Kaufman, & Ponitz, 2009; Hamre & Pianta, 2005; Rimm-Kaufman et al., 2002; Walkington & Marder, 2014). The findings that observation scores are related to student outcomes differently across different groups of students are not consistent, however. For example, CLASS scores appear to be equally predictive of outcomes for Hispanic and non-Hispanic students and students whose first language is and is not English in pre-school (Downer et al., 2012). These discrepancies could stem from CLASS being more robust to the instructional needs of different groups of students or to the fact that differential validity of observation instruments only occurs in certain grades. Thus, there is a need for further exploration of the differential validity of observation scores across groups of students. Assuming the effects of observed teaching quality on student outcomes replicate, research needs to examine whether such interaction effects are driven by different types of students needing different types of instruction, or by the unique sensitivity of disadvantaged students to poor instruction, or perhaps by other factors not yet identified.

III.2.3.2. Subject Observed teaching quality scores may vary across subjects (e.g. math, English, science, social studies) because teaching approaches differ across subjects (J. J. Cohen, 2015b; Stodolsky, 1984). The evidence of subject differences in scores on the observation instruments studied here is mixed, with some studies finding higher scores in English than math (Chaplin et al., 2014) while others find no differences (Curby et al., 2011; Pianta & Hamre, 2009). Only one study that I could find explored within-teacher differences in instructional quality across subject (Curby et al., 2011). It found math and English instruction had fewer positive emotions, were more controlling, and were more productive than instruction on average. Various effects were found for other subjects most of which were small. Importantly, the effects on observed teaching quality varied across grades in complex ways, suggesting subject effects are grade specific. Overall, then, more research is needed to explore how and why subject affects observed teaching quality with a focus on grade level moderators and whether effects are between or within teachers.

There is more consistency in research that has shown the relationship of observation scores and VA scores varies across subjects. This relationship was found to be stronger in math than English (Chaplin et al., 2014; Kane et al., 2012), which likely reflects the greater role of non-school forces in the learning of English compared to math. Interestingly, this difference was not seen for the higher-order English test used in the MET study, which tested mostly writing performance, a skill that is apparently less affected by non-school forces (Kane et al., 2012). Thus, not only might the subject being taught affect observed teaching quality, but it might impact the relationship between teaching quality and teacher quality.

III.2.3.3. Grade Levels The grade level of students being taught also has been found to have a large effect on observed teaching quality in previous research. The MET project, for example, found large grade effects, with middle school teachers (grades 6-8) scoring significantly lower than elementary school teachers (grades 4-5) (Grossman et al., 2014; Mihaly & McCaffrey, 2014). Studies of teacher evaluation systems have found similar effects of grade level on observed teaching quality (Chaplin et al., 2014). Importantly, this effect was not explained by differing teacher or student characteristics across grades, but rather seemed to reflect differences in curriculum and student maturation effects (Mihaly & McCaffrey, 2014; Walkington & Marder, 2014). However, as in research on subject effects on teacher quality, the lack of studies that observed the same teacher teaching more than one grade makes any conclusions as to the cause of grade differences in teaching quality unclear. The effects could be driven by teacher sorting to preferred grades.

There is also mixed evidence concerning the extent to which the relationship of observed teaching quality scores and VA scores varies across grades. Some studies found the correlation of observation scores and VA scores was higher in elementary grades (Chaplin et al., 2014), some found the correlation was higher in middle grades (Walkington & Marder, 2014), and others found no difference across grades (Mihaly & McCaffrey, 2014). Walkington and Marder (2014) looked in detail at why the relationship between an observation score and VA score was higher in middle schools using the UTeach Observation Protocol and MET data, finding that student behavior and school climate often affected VA scores but had little effect on observed teaching quality. They also identified content as an explanation of the varying relationship between observation and VA scores because the focus on lower-order tasks in instruction detracted from observed scores, but not VA scores. Thus, the differential validity of observation scores across grades appears driven by observation scores responding to aspects of teaching quality unrelated to VA scores. This suggests that the differential validity of observation scores across grades is likely instrument and test dependent.

III.2.3.4. Schools and/or Districts The goal of this thesis is to explore the relationship between observed teaching quality and "true" teacher quality. Schools and districts may have an important moderating effect on this relationship (Blazar, Litke, & Barmore, 2016; Jiang & Sporte, 2016; Lynch, Chin, & Blazar, 2015), though this potential effect has been ignored in most past work. Schools may affect the relationship between teaching quality and true teacher quality because, as many have asserted, schools have distinct instructional cultures that affect teaching practices (Bryk, et al., 2010; Ladson-Billings, 2008), though convincing evidence demonstrating how schools affect observed teaching quality is harder to come by. Cohen and Brown (2016) found that observation scores and VA scores are unrelated in schools with positive school environment ratings but positively related when the school has negative school environment ratings, suggesting that school environment moderates the validity of observation scores. This study, however, was conducted with a small sample and tables presented in the paper show one outlier teacher whose data could be driving the moderating effects of school environment. Cohen and Grossman (2016) similarly reported that the relationship between observation scores and VA scores varied across schools, but it is unclear whether the two papers shared common schools. The authors of the two papers suggested that schools with a positive environment had more shared responsibility for student learning so the instructional skill of the classroom teacher was less connected to student learning. Holtzapple (2003) found similar moderation effects of school climate, which were driven by higher VA scores for the worst teachers in good schools compared to the VA scores of the worst teachers in bad schools. If the power of observation scores to predict VA scores is only in the lower tail of the distribution (e.g. Holtzapple, 2003), this could explain the results from Cohen and colleagues (2016; 2016) without appealing to moderating effects of schools.

In general, however, much more work is needed to understand how school (or district) environments might affect observed teaching quality and how this relationship affects the extrapolation to true teacher quality. This is vital because teacher evaluation systems are extrapolating across schools to make between-school comparisons of teachers, which rely on estimating the causal impact of the teacher on teaching quality. If schools affect a teacher's ability to provide instruction, the relationship of teaching quality and teacher quality will vary across schools, leading school effects to contaminate the estimates of teacher quality (Gitomer & Bell, 2013).

III.2.4. Summary of Facet Effects on Observation Scores Current uses of observation instruments often focus on questions that require causal attributions to teachers (Bell et al., 2012; Gitomer & Bell, 2013). We wish to know whether one teacher meets a given threshold of quality or performs better than another teacher. This is a challenging problem because observed teaching quality varies widely with characteristics of the lesson and classroom being observed. Teaching is a situated task and must be understood as such if we want to make appropriate conclusions about teacher quality (J. J. Cohen & Goldhaber, 2016; Kennedy, 2010). In this section, I reviewed the evidence that currently exists about factors that are systematically related to differences in average scores on observation instruments. Unfortunately, this evidence is often inconsistent or incomplete across studies. Further, it is often unclear if effects generalize across the single instrument used in the study.

However, it is clear that three broad classes of hidden facets affect observed scores on observation instruments. These hidden facets—labeled here as System Design facets, Curriculum and Instruction facets, and School Organization facets all represent groups of variables that have been shown, with varying amounts of replication and consistency, to affect observed teaching quality. While these classes of facets affect averaged observed teaching quality, much less is known about how these hidden facets affect the measurement properties of observation scores—the reliability of estimates, bias in estimates, and the validity of score estimates. In this thesis, I hope to provide information on this point by incorporating these hidden facets into a broader measurement framework and statistical models that examine how observed scores generalize across facets of measurement.

III.3. Validity of Classroom Observation Scores

Up to this point, I have reviewed past work relevant to my first two research questions which explore the effect of the planned and hidden facets of measurement on observed teaching quality. In this section, I turn towards evidence supporting the validity of score estimates derived from classroom observation instruments. The majority of research that has explored the validity of using observed teaching quality to make conclusions about teacher quality has focused on the relationship between classroom observations and VA scores, though some work has connected observations to teacher knowledge and student survey measures of teacher quality.

The evidence linking observation scores and VA scores is growing (e.g. Kane et al., 2013; Milanowski, 2011; Schacter & Thum, 2004). There are, however, concerns about this evidence. The relationship between observation scores and VA scores may be driven mostly by effects in the lower tail of the distribution (Grossman, Cohen, Ronfeldt, & Brown, 2014; Holtzapple, 2003; Lynch et al., 2015). Further, the non-random sorting of students to teachers raises the potential that shared measurement error is driving this relationship. Using MET randomization data, Garrett and Steinberg (2015) found that observation scores can causally identify effective teachers only in math (and not English), providing only partial support for the claim that the relationship between observation scores and VA scores is not the result of shared error stemming from the non-random sorting of students. The lack of an identifiable causal relationship stems, in part, from the weak connection between observation scores and VA scores and high levels of measurement error in both measures. This weak connection is unsurprising given that different approaches can lead to student learning, students may respond differently to the same instruction, learning occurs outside of classrooms, and the long history of low correlations from process-product work (Croninger & Valli, 2009; Good, 1979; Muijs, 2006; Seidel & Shavelson, 2007). Observation scores have also broadly been connected to other outcomes, such as teacher knowledge (Bell et al., 2012; Hill, Ball, Blunk, Goffney, & Rowan, 2007) and the quality of student work (Matsumura, Garnier, Slater, & Boston, 2008). Thus, there is broad and consistent evidence that supports the validity of observation scores, at least correlationally, but the relationships are weaker than desired.

This evidence of the validity of observation scores does not directly address the question of whether hidden facets affect the validity of observation scores. Only the MET study directly addressed this, but MET research on the issue was done in passing and the researchers simply noted a correlation above 0.9 between teacher score estimates before and after adjusting for the facets of student characteristics, implying student characteristics did not have an effect on the validity of observation scores (Kane et al., 2013). It is not clear how well this finding generalizes though because, as I discussed before, other studies suggested more meaningful changes to scores as a result of adjusting for student characteristics (e.g. Whitehurst et al., 2014).

III.4. Chapter Summary

In this chapter, I reviewed the past research that explored the connection between the contexts of measurement and observed teaching quality. Here, I summarize this work, connecting it directly to my research questions. This research shows clearly that observed teaching quality varies, over both planned error facets and contextual features of lessons not generally considered in measurement models (i.e. hidden facets). The various planned error facets make large contributions to observed score variance and these contributions are fairly consistent across studies, though no evidence exists regarding how accurately they are estimated.

There is less evidence for consistency in the research on hidden facets, however. The research here is shallow, contains few replications of any given result, and is splintered such that most results apply to only a single classroom observation instrument. The one exception to this is the strong connection between observed teaching quality and student prior achievement and demographic characteristics, which has been robustly shown across many studies and instruments. For most hidden facets, though, more research is necessary to explore the generalizability of findings across samples and instruments. Given this, very little is known that directly relates to my second research question.

Few studies have explored how adjusting for hidden facets might change estimates of teacher quality or examined the implications of these changes for issues of validity. Further, little research is explicit about whether facets act within or between-teachers, which, as I have argued, is important for understanding their effect on teacher's score estimates and for understanding when extrapolation is necessary to compare scores across teachers. To be sure, there is growing interest in adjusting observation scores for contextual features, but this interest has largely focused on adjusting measures for student characteristics and has not connected these adjustments to a broader measurement framework that fully addresses the many questions associated with using adjusted or unadjusted models for score estimation.

Research on the validity of observation scores has focused almost solely on connecting observed teacher score estimates to other measures of teacher quality. The validity research has yet to explore whether adjusting for hidden facets may help improve the validity of estimates of teacher quality. Further, there has been little explicit exploration of how much the relationship between scores from observation instruments and other measures of teacher quality varies across levels of hidden facets, though some studies have shown the correlation of observation scores and VA scores varies across schools, subjects, and grades. Overall, then, past research provides support for my claims that the context of measurement matters in understanding observed teaching quality, but largely leaves open the implications of this relationship, especially the question of how contexts of measurement might impact estimates of observed teacher quality.

Chapter IV. Methods

In this chapter, I review the data sources used in this thesis and describe my analytic models in detail. I start by describing the UTQ study, the classroom observation instruments, including a short instructional log, and briefly discuss the value-added scores used by the UTQ study. I then turn to describing the GTheory models that I estimated to test the research questions, highlighting the models used to examine each question.

IV.1. Understanding Teacher Quality (UTQ)

The data for this thesis were collected as part of the Understanding Teaching Quality project (UTQ; http://utqstudy.org/). This project was designed to examine how well existing teacher observation instruments measured teaching quality with the goal of increasing the value of these tools for personnel evaluation and instructional improvement. The UTQ project conducted live and video observations of mathematics and English language arts teachers in grades 6-8 in three large school systems in the southeastern United States from 2009-2011. The project had a sample of 458 volunteer teachers (228 of whom taught English), with roughly half the teachers in the project participating in each of the two school years when research was conducted. The data reported in this thesis focuses only on ELA classrooms because the PLATO protocol included an instructional log that allows me to study some of the hidden facets discussed in the last chapter directly. In the single year they participated, each teacher was observed and videotaped teaching one lesson on four separate days across 2 sections[16]. Each teacher's instruction was scored using the CLASS, FFT, and PLATO instruments.

In UTQ, a total of twelve raters participated in scoring[17]. These raters received multiple days of training on each of the three observation instruments under study here: CLASS, FFT, and PLATO. All raters were certified by observation protocol developers to conduct scoring of the relevant instrument before beginning to score lessons according to rules developed by the observation protocol developers. In addition, calibration exercises were conducted every 3 weeks during the course of the study to maintain reliable scoring over time[18]. Calibration consisted of scoring a video with master codes, discussing the video and scores, and receiving feedback. No actions were taken when observers did not score accurately during calibration exercises.

As I discussed earlier in this thesis, the assignment of raters to scoring is important to interpreting rater effects, so I will spend some time here describing the process of assigning raters to lessons and the organization of the scoring process. In UTQ, there were two phases of scoring. The first phase included live scoring of year 1 teachers for 90% of the year 1 lessons. Live scoring only happened for one instrument per day, so only 30% of year 1 ELA videos have live scores on each individual instrument (CLASS, FFT, and PLATO). Phase 2 scoring began after the end of phase 1 and consisted of video scoring both years of lessons. Videos were randomly assigned to raters, thus randomizing both the rater scoring the video and the order in which videos were scored, though year 1 videos were scored, on average, earlier than year 2 videos because year 2 data collection was ongoing during scoring. In UTQ, raters were assigned to score at most one video per teacher, though assignment of live

[16] Four classrooms were observed only once due to scheduling problems.

[17] One rater only completed the live scoring.

[18] Observers completed calibration on one of the three instruments that they were scoring each week.

and double scoring was done independently of assigning primary scoring tasks (note that no raters were assigned to a day they previously scored in person). Double scoring was completed for one of the four videos submitted by each teacher (25% rate) and conducted by a randomly assigned rater. The combination of live and double scoring resulted in about 60% of videos scored by one rater, 34% by two raters, and 5% by three raters, allowing 39% of videos to contribute to my estimate of the rater-by-day variance component (which requires two raters to score the same day of instruction). There were 471 cases of raters scoring two videos from the same teacher (174 unique teachers) and 34 cases of a rater scoring three videos from the same teacher (33 unique teachers). These cases form the basis of rater-by-teacher estimates (which requires the same rater to score multiple days of instruction from the same teacher). This scoring setup provides a large minority of videos with multiple raters to estimate inter-rater reliability and ensures many raters view each teacher; but the setup limits the number of times a single rater scores days of instruction from any given teacher, restricting the amount of data available to estimate some rater biases, such as rater biases against specific teachers.

All three instruments under study divided the day of observation into occasions based on instrument protocol (described below) and assigned scores for each occasion sequentially. Days of instruction consisted of between 1 and 7 occasions for PLATO and CLASS. FFT, though, used 30 minute occasions so few videos have more than one occasion and none have more than two. Live observations contained a time gap between scoring occasions equal in length to the instrument's scoring period. The scores from videos contained no such time gap between occasions. Thus, beyond the first occasion of a lesson, the video and live observations were scored on somewhat different time periods. When scoring both types of videos, observers started at the beginning of the lesson and progressed sequentially through scoring all occasions. While observers were instructed to score each occasion independently, this scoring design has the potential to reduce the variance between occasions because of carry-over effects (i.e. the scores for the first occasion affect scores for future occasions; see Ho & Kane, 2013). However, this approach to scoring makes the video scoring process similar to the live scoring process, hopefully minimizing the effect of scoring mode.

IV.2. Observation Instruments

My thesis focuses on the three observation instruments used to score English classrooms: CLASS, FFT, and PLATO. Using three instruments allowed me to more fully characterize the nature of each day of instruction than would be possible using only a single instrument. Further, by comparing the effects of particular hidden facets across instruments, I can explore instrument biases.

IV.2.1. Classroom Assessment and Scoring System-Secondary (CLASS)

The Classroom Assessment and Scoring System-Secondary (CLASS); Pianta et al., 2007) was developed as an extension of a project examining the impact of classroom quality on child development outcomes under the premise that the proximal interactions in the classroom will lead directly to these outcomes. CLASS purports to be content neutral, focusing on the interactions between teachers and students. It was originally created and used across the first half of a school day (in grades K-3), capturing unstructured time between lessons. In fact, CLASS developers typically encourage the observation of unstructured time between classes. When used in higher grades, however, it has focused only on instructional periods, much like the other instruments.

There are three broad measurement domains in CLASS, and these domains are broken into 11 dimensions. The domain of Emotional Support focuses on the emotional and social tone of the classroom, largely growing out of the literature on attachment theory and self-determination theory (Pianta & Hamre, 2009). It is composed of four dimensions: Positive Climate, Negative Climate, Teacher Sensitivity, and Regard for Adolescent Perspective.

Negative Climate is unique in that higher scores denote lower quality. In this thesis, I reverse code Negative Climate so higher scores capture higher quality. The domain of Classroom Organization captures the efficiency and management of the classroom. It is comprised of three dimensions: Behavior Management, Productivity, and Instructional Learning Formats. The domain of Instructional Support captures the nature of instructional interactions between students and the teacher, focusing on the development of higher order thinking skills. It is composed of four dimensions: Content Understanding, Analysis and Problem Solving, and Quality of Feedback. Last, CLASS codes student's engagement, which is viewed as an outcome measure.

In the UTQ study, each day of observation was divided into 15 minute occasions with each occasion scored independently. When live scored, the lesson was divided into 22 minute occasions with 15 minutes spent observing the lesson followed by 7 minutes of scoring. When scored from video, raters paused the video for 7 minutes to score an occasion, leaving no breaks between 15 minute scoring occasions. Occasions less than 10 minutes in length were not scored. This led to 5% of videos with 2 occasions, 65% with three occasions, and 30% with 4 or more occasions scored. Twelve raters scored CLASS live and 11 scored videos using CLASS. Raters received multiple days of training prior to scoring and passed a certification test that required them to score 80% of dimensions within one point (on 7 point scale) of a previously determined master score across 5 test videos.

IV.2.2. The Framework for Teaching (FFT) The Framework for Teaching (FFT) was developed as an extension of the work developing the Praxis III observation system for teacher certification (Danielson, 2000). It too purports to be content neutral, adopts a constructivist view of student learning in principle, and includes items that measure teacher preparation and planning and teacher professional responsibilities, as well as instruction. While most teacher evaluations systems that have adopted FFT use all the components of the instrument, at least in a modified form, research on teaching has focused mainly on Domain 2 (labelled the Classroom Environment) and Domain 3 (labelled Instruction). These are the two domains that can be scored solely from observations of classroom instruction. The Classroom Environment domain is comprised of five dimensions: Creating an Environment of Respect and Rapport, Establishing a Culture of Learning, Managing Classroom Procedures, Managing Student Behavior, and Organizing Physical Space. The domain of instruction is also comprised of five dimensions: Communicating with Students, Using Questioning and Discussion Techniques, Engaging Students in Learning, Using Assessment in Instruction, and Demonstrating Flexibility and Responsiveness. In UTQ, one dimension of Domain 1, Planning and Preparation was also scored: Demonstrating Knowledge of Content and Pedagogy.

FFT had its own unique scoring design that differentiated between live observations and videos. Live observations were scored by observing instruction on a given day for 30 minutes and then scoring for 15 minutes. This was repeated if time allowed. Occasions had to be longer than 20 minutes to be scored. Video records were scored in 30 minute occasions with no time lapses between occasions. This led to most videos having only one occasion and 12% with two occasions. Twelve raters scored FFT live and 11 raters scored videos. Raters received training and passed a certification test the required them to score 50% of dimensions exactly (on a four point scale) and less than 25% of dimensions 2 or more points from the master score across 4 videos.

IV.2.3. Protocol for Language Arts Teaching Observations (PLATO)

The Protocol for Language Arts Teaching Observations (PLATO) (Grossman et al., 2013) was developed as an extension of the CLASS instrument to focus specifically on English language arts instruction. Like CLASS, PLATO is intended to focus on interactions in the classroom under the assumption that proximal interactions cause student learning. Unlike CLASS, however, PLATO explicitly examines content specific (i.e. English) instructional practices because PLATO developers believe that content and teaching cannot be separated and that best practices differ across subject areas (like English and math). Since its initial development, PLATO has become an independent observation protocol by dropping dimensions from the CLASS instrument, such as the Emotional Climate items. Currently, four major domains are measured by PLATO: Disciplinary Demand of Classroom Talk & Activity, Contextualizing and Representing Content, Instructional Scaffolding, and Classroom Environment. These domains are broken down into 13 dimensions in the version of PLATO used for UTQ: Purpose (expressed clarity of the lesson), Intellectual Challenge, Representation of Content (teachers' ability to represent content to students through effective and meaningful explanations), Connections to Prior Academic Knowledge, Connections to Personal and Cultural Experiences, Models/Modeling, Explicit Strategy Instruction, Guided Practice, Classroom Discourse, Text-Based Instruction (the presence and use of texts during class), Accommodations for Language Learning (ways teacher incorporates strategies for English language learners), Behavior Management, and Time Management.

In the UTQ study, PLATO was scored by breaking each recorded lesson into 15 minute occasions with each occasion scored independently. When live scored, the lesson was divided into 23 minute occasions with 15 minutes spent observing the lesson then 8 minutes given over to completing the scoring task. When scored from video, raters paused the video for 8 minutes to score the lesson, leaving no breaks between 15 minute scoring occasions. Occasions less than 10 minutes in length were not scored. This led to 5% of days with 2 occasions, 65% with 3 occasions, and 30% with 4 or more occasions. Unlike CLASS and FFT, only 6 raters scored PLATO (one rater left after live scoring leaving 5 to score videos). Raters received multiple days of training prior to scoring and passed a certification test that required them to score 80% of dimensions correctly across 5 videos.

IV.3. PLATO Log

The PLATO log--which is a separable component of the PLATO observation instrument--is an important additional instrument studied in this thesis. It serves as a unique data source to record what would otherwise be hidden facets of measurement. In particular, the PLATO Log has two main parts. The first records the content domain of the lesson for each scoring occasion (Reading; Writing; Literature; Oral Language; Vocabulary/Word Study; Grammar/Spelling; Research Strategies) and whether each content domain was a major focus, minor focus, touched on briefly, or not touched on during instruction. Importantly, raters can select multiple content domains for a given occasion. When lessons had a major focus on reading, writing, or literature, an additional section of checklist items was completed. I did not use this additional section of the log, however, due to concerns about missing data and rater error. In order to examine the hidden facet of content domain, I aggregated Content Domain variables to the day-level so variables could be used with FFT. I scored a day of instruction as having a sustained focus on a specific content domain if that domain was a major or minor focus for two consecutive occasions on a given day. This operationalization balanced identifying lessons with a strong focus on a given content domain and obtaining enough lessons within each domain for stable estimates of the effect of the content domain. I dropped Oral Language, Vocabulary, and Research Strategies because too few days had these as a sustained focus. These coding decisions were all made prior to testing effects on content domain on observed teaching quality. Of the 901 total days observed, 74 (8%) had a sustained focus on reading, 203 (23%) had a sustained focus on literature, 234 (26%) had a sustained focus on writing, 235 (26%) had a sustained focus on grammar, and 240 (27%) had no sustained focus. The 240 days with no sustained focus generally shifted between a focus on multiple content domains or focused on excluded domains, though 80 days contained no major or minor focus on any content domain. Most teachers were observed across different domains on different day with fewer than 15 teachers submitting three or four days on the same content domain. Most days of instruction had a sustained focus on only one content domain.

The second section of the log was a set of checklist items that denoted classroom activity structure[19]. I used this checklist to create a measure of the construct of interaction structure. After examining this data, I aggregated items to the day-level using the same concept of sustained focus. When an activity structure (e.g. teacher talk/lecture) was indicated as present for two consecutive occasions, I scored the lesson has having a sustained focus on that activity structure. I used these lesson-level variables to construct three new variables: recitation/lecture (a sustained focus on teacher talk/lecture OR short student response OR student presentations); discussion (a sustained focus on small group discussion either structured or unstructured OR whole class discussion); and independent work (a sustained focus on either independent work OR independent reading from the reading domain). These three composite variables capture structures organizing student/teacher interactions, which I termed "interaction structure". Interactions form the basis for scoring lesson quality. Thus, the structure used to organize instructional interactions is an important factor to consider when exploring the effect of day characteristics on observation scores and when testing for instrument biases. For example, in recitation/lectures, the interactions focus on listening and responding in limited, controlled ways. In discussions, interactions are more free and open, based on the topic of discussion. In independent work, limited interactions occur. There is overlap in these categories and none is operationalized perfectly, but my approach to measurement should provide a broad sense of how the structure of instructional

[19] The full range of possible classroom activity structure items is Teacher Talk/Lecture; Short student responses to teacher questions; Small group/partner discussions unstructured; Small group/partner discussions structured (literature circles, etc); Whole group discussion; Student presentations; Independent work; Teacher (or student) uses students' primary language to introduce or explain key concepts, terms, etc; Teacher provides differentiated assignments or assessments.

interactions affects observation scores. The goal here is to provide evidence to guide future research rather than to definitively address research questions. There were 659 days (73%) with a sustained focus on recitation/lecture, 454 days (50%) with a sustained focus on discussion, 92 days (10%) with a sustained focus on independent work, and 84 days (9%) with no sustained focus on any interaction structure. Many days had an interaction structure in two or more areas: 304 (34%) days had a focus on discussions and recitation/lecture, 75 (8%) days focused on recitation/lecture and independent work, and 39 (4%) had a sustained focus on both discussions and independent work.

Because of the importance of the PLATO log to my study of hidden facets, a number of limitations of the instrument should be noted now. The inter-rater reliability is low, especially for the interaction structures. Table 4.2 shows inter-rater reliability statistics for the log items. The left column specifies the facet being described and columns show, in order, the statistics of percent raw agreement, Cohen's Kappa, Negative Agreement, and Positive Agreement (Gwet, 2012). Negative Agreement captures agreement conditional on either rater coding the variable as 0 and Positive Agreement captures agreement conditional on either rater coding the variable as 1. The Kappa values are below 0.66 for all content domain items and close to zero for interaction structure. Recommended minimum Kappa values are 0.6 with 0.8 preferred (Graham et al., 2012). Despite the low Kappa values, the agreement rates for content domain were comparable to gateway items in the Study of Instructional Improvement (SII) log while the agreement for interaction structure was only slightly below the agreement on back-end items of the SII log (Camburn & Barnes, 2004; Rowan & Correnti, 2009). The lower than desired reliabilities may have stemmed, in part, from the PLATO log not being part of the rater certification and calibration process, which may have led to a reduced focus on learning to score these items well in training.

Table 4.2: Inter-Rater Reliability Statistics for the PLATO Log variables

Facet	Percent Agreement	Kappa	Negative Agreement	Positive Agreement
Content Domain				
Reading	92%	0.40	96%	44%
Writing	87%	0.60	92%	68%
Literature	84%	0.52	90%	62%
Grammar	87%	0.66	91%	75%
Interaction Structure				
Recitation/Lecture	50%	-0.04	35%	59%
Discussion	55%	0.04	64%	38%
Independent Work	88%	-0.03	94%	04%

Note. Percent Agreement =Percent Raw Agreement; Kappa=Cohen's Kappa; Negative Agreement =Negative Percent Agreement; Positive Agreement =Positive Percent Agreement.

Another limitation of the log is the limited set of variables capturing instruction. A complete exploration of hidden facets would require a much wider set of variables, including instructional goal (i.e. review, introduce new material, independent practice), grouping structure (i.e. small group, whole class), the cognitive demand of the content taught, and other lesson characteristics. The structure of breaking days down into 15 minute observation intervals also works against the effectiveness of the log. Fifteen minutes is an arbitrary length of time that may not capture natural phases of instruction, as I have discussed.

IV.4. Value-Added Scores

In studying the validity of the CLASS, FFT, and PLATO instruments, I will be correlating the scores teachers received on these instruments to teachers' Value-Added (VA) scores as calculated by UTQ authors. I only briefly discuss these scores and do not delve into the specific statistical models used to create the scores (see Lockwood & McCaffrey, 2014 for a detailed discussion). To begin this discussion, please note that I used one of the provided VA scores from the data set (called "lr6 estimates" in the data set). In my view, there are important challenges associated with correlating these VA scores to UTQ observation scores. UTQ provides two (lr6) VA scores, the same-year scores and an alternate-year scores (i.e. the teacher's VA score from the previous year). Generally, the alternate-year VA score is preferred because the same students do not contribute to observed teaching quality and the alternate-year VA score. However, the alternate-year scores are correlated with current year students' prior achievement. That is, the gains made by last year's students are correlated with the incoming ability of this year's students. This is a common, though rarely discussed, source of bias to VA scores, which the UTQ models have minimized, but not eliminated (Lockwood & McCaffrey, 2014). Moreover, this correlation is troubling because classroom observation scores are also correlated with students' prior achievement. Thus, the relationship between alternate-year VA scores and current students' prior achievement is an indication of (potentially) shared measurement error between observation scores and alternate-year VA scores. This shared measurement error will bias correlations between the two measures[20]. The current year VA scores, due to explicit controls for prior achievement, do not contain this particular source of correlated error, but likely contain other unknown correlated errors stemming from the fact that the same students contribute to both measures. My solution to this problem in the analyses described below is to control, within a regression framework, for the prior ability and the demographics of current students so the estimated relationship between observation scores and VA scores is independent of the prior achievement and student demographics effect. This assumes that students' prior achievement and student demographics capture all of the correlated measurement error, which may not be true.

IV.5. Generalizability Theory Analytic Models

Having discussed the UTQ study procedures and the variables I will be analyzing. I now lay out the analyses that I conducted for this thesis. I start by introducing the base GTheory model that I estimated. Here, I focus on the specific model that I ran, highlighting

[20] I have never seen this discussed and it seems to be a major threat to much research currently going on (see Kane et al., 2011, Kane et al. 2012...) that argue for using correlations with alternate year VA scores to avoid biases.

how it varies from the previously introduced full model (Equation 1). I then provide an overview of the "Decision Studies" I used to calculate score reliability from the GTheory model. Next, I introduce more complex models that control for the three classes of hidden facets. These models, to varying degrees, embrace the situated nature of teaching. Comparisons of these models to the Base model allow me to explore the impact of controlling for hidden facets on the measurement properties of observation scores and on estimates of teacher scores. In the next section, I describe my approaches for comparing model estimates, highlighting the implications of these comparisons. Additionally, I describe how I will examine the validity of teacher estimates across models, including differential validity across facets.

IV.5.1. Base Model (Base) I began building my GTheory statistical model from the full model described earlier (Equation 1). The model in Equation (1) is highly complex and contains many terms that are not well separated, which stems from the partial crossing of raters and days in the UTQ design[21]. For example, in UTQ data, the item-by-rater-by-day ($\upsilon_{ir(d:s:t)}$) and the item-by-day ($\upsilon_{i(d:s:t)}$) terms in Equation (1) are only distinguishable on the minority of days that have multiple raters (usually 2 raters). I fit the statistical models reported here using Restricted Maximum Likelihood (REML) with the package $lme4$ (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2016a), which should be capable of estimating both facets despite this challenge. However, the inclusion of all facets still led to challenges, such as lack of convergence and extremely long model run times (which is especially problematic given the bootstrapped confidence intervals for the estimated variance components). Further, there is a difficulty in interpreting the facets with high-level interactions as distinct from the pure residual variance. These considerations led me to

[21] The correlation between the day and rater-by-day variance component estimates across the bootstrapped samples was close to -0.8.

combine the three-way interactions and the rater-by-occasion term into a single residual term in the models presented in this thesis[22]. After combining these facets, I ran the models, finding close to zero variance at the section-level across all instruments[23]. In order to save computational time, I therefore eliminated section-level facets from the model. Last, in consideration of the fact that each instrument has a fixed set of items, I chose to include items as a fixed effect. The interactions of items with other facets are still modeled as random effects, which introduces some minor error to the model as it assumes that all items vary across occasions, days, and teachers to the same extent (i.e. it calculates average item variation over different facets). My Base model, then, is below:

$$\begin{aligned} \mathrm{X}_{\{ir(o:d:s:t)\}} &= \mu + \upsilon_t + \upsilon_{\{d:s:t\}} + \upsilon_{\{o:d:s:t\}} \\ &\quad + \upsilon_{\{it\}} + \upsilon_{\{i(d:s:t)\}} + \upsilon_{\{i(o:d:s:t)\}} \\ &\quad + \upsilon_r + \upsilon_{\{rt\}} + \upsilon_{\{r(d:s:t)\}} \\ &\quad + \upsilon_{\{ir\}} + \epsilon_{ir(o:d:s:t)} \\ \epsilon_{ir(o:d:s:t)} &= \upsilon_{\{r(o:d:s:t)\}} + \upsilon_{\{irt\}} + \upsilon_{\{ir(d:s:t)\}} + \upsilon_{\{ir(o:d:s:t)\}} \\ \mu &= \beta_i \end{aligned} \tag{Base}$$

where $\epsilon_{ir(o:d:s:t)}$ represents the residual, μ is a stand-in for all fixed effects, β_i is item fixed effects, and all other variables are as before (see equation 1). Teacher quality, as estimated from this model, represents the teacher's average enacted teaching quality over the time period from which days are sampled (usually a year).

A few notes of caution about interpreting parameters in this model are necessary. I modeled the outcomes as continuous, but they are actually ordinal. There are two problems with this. First, only a limited number of item scores are possible (i.e. CLASS is a 7 point

[22] Models with the rater-by-occasion term generally had slightly lower occasion facet variances and small rater-by-occasion facets. Suggesting, at least when raters are assigned at the day-level, few rater-by-occasion effects. Further, the item-by-rater-by-teacher and item-by-rater-by-section effects were generally small. The item-by-rater-by-day facet was generally large, almost equal in size to the residual term, but it is hard to interpret this facet apart from being a residual error (the residual is the item-by-rater-by-occasion facet).

[23] It is difficult to conclude why the section-level variance is so low, especially because the effect of student composition appears to be quite large across a range of studies (including in the UTQ data). It is possible that the model simply does not have enough days of instruction per section to estimate a stable score for sections. After all, the reliability of teacher variances is quite low, much less for sections, which have half as much data supporting their estimation.

scale; FFT/PLATO are 4 point scales). Modeling scores as continuous is common practice in GTheory, though the limited range of outcome values can affect variance component estimates (Shavelson et al., 1986; Shumate, Surles, Johnson, & Penny, 2007). Prior GTheory work examining the impact of assuming a continuous outcome has been conducted on very simple models (compared to this one) so past work may not generalize to this model. Nonetheless, I have run models in which I first averaged across occasions or items, increasing the range of values the dependent variable ($X_{ir(o:d:s:t)}$) may take, as a sensitivity analysis. These alternative models lead to very similar conclusions as the models presented in this thesis, though they provide less ability to explore error facets.

A second problem stemming from the ordinal nature of the outcome is that the distance between rubric score points in my model is assumed constant (i.e. $X_{ir(o:d:s:t)} \in \{1-4\}$ for PLATO/FFT). Because the data is ordinal, there is no reason to prefer using 1-4 as compared to 1, 3, 4, and 9 or any other set of increasing numbers. To test if the choice of using 1-4 had an effect, I used Correspondence Analysis (Greenacre, 2005) to rescale item responses[24]. Using the re-scaled values, I ran the same GTheory model described above (Equation Base) as a sensitivity test. This alternative approach led to approximately the same percentage of variance across each facet as the model using the original values, suggesting robustness to the equal-interval assumption. All results presented use the original item responses (i.e. $X_{ir(o:d:s:t)} \in \{1-4\}$ for PLATO/FFT). I could have run these models using an ordinal link function, but this would eliminate the residual variance, preventing a full analysis of variance across facets. Additionally, ordinal models are non-linear and significantly harder

[24] Correspondence analysis is akin to an ordinal principal components analysis. Item scale scores (i.e. the numbers 1-4) are re-scaled to maximize the correlation between individual dimension scores and the average score across the items. This leads to the rescaling that maximizes the percentage of variance explained by the average score across items. For example, the scale points of 1-4 on FFT's Respect and Rapport item were changed so that 1 => -4.2, 2 => -1.9, 3 => 0.4, and 4 => 3.3. This is admittedly a somewhat arbitrary way of rescaling the data, but it provides a sensitivity test for the specific values used as scale points.

to fit. There is no software I could find available that would fit this complex of a model[25], in fact the algorithms necessary to do so are an active area of research (Schilling & Rowan, Personal Communication).

Another potential challenge with this model is the assumption of equal residual variance across items, because the unconditional variance of item scores varies across items. I fit a model allowing for this heterogeneity [which I did in a Bayesian framework using $brms$ (Bürkner, in press). This model resulted in estimated variance components that were not significantly different from those in the $lme4$ model, suggesting robustness to heterogeneity of variances. Last, there is the threat of auto-correlation across occasions within a lesson, which I did not model using $lme4$. I did fit models (using $brms$ in a Bayesian framework) that included this auto-correlation and these models showed no significant auto-correlation across occasions.

A common flaw in GTheory applications in research on teaching is the lack of reported uncertainties in the estimated variance parameters. This occurs, no doubt, because most common statistical software does not provide uncertainty estimates for variance components and because of the computational demands of creating these uncertainty estimates. This is unfortunate because the data structures are highly complex with many levels of nesting and partially-crossed data, which may lead to high levels of uncertainty in model estimates. In the results reported here, I use fully parametric bootstrapping (Brennan, 2001) to generate 95% confidence intervals for variance component estimates using the percentile bootstrap (Hesterberg, et al, 2005; Efron & Tibshirani, 1994; see Appendix G for a deeper discussion of this bootstrapping approach and comparison with alternative methods). Here, I simply state that while percentile bootstrap methods are often found to be too narrow,

[25] If I moved to a Bayesian framework, I could, in theory, find software to do this fit. In fact, I tried this (briefly) using $brms$ to call $STAN$, but could not get the chains to mix.

they are consistent with error O(n^-0.5) under many conditions (Hesterberg, et al, 2005; Efron & Tibshirani, 1994)[26]. It is important to note what these confidence intervals do and do not represent. Parametric bootstrapping assumes the estimated model matches the population model and re-samples new data under this assumption. This gives the sampling variation of parameter estimates, but only under the assumption the original model is (approximately) correct. This is a challenge here because, as I have noted, the structure of the data may make it difficult to find large variance estimates for some parameters (e.g. section variance, rater-by-teacher variance). If this is the case, the confidence intervals may not be accurate.

The base model above (Equation Base) will form the foundation for analyses that address the first research question. These analyses focus on examining the relative variances associated with different planned facets (e.g. $var(\upsilon_r)$) and the uncertainty of these estimates. Comparisons of variance components, especially across the three instruments, can elucidate which facets of measurement most affect observed scores. For example, if rater-by-teacher variance is low, there is a limited amount of rater-specific bias based on teacher characteristics[27]. Again, however, the reader is cautioned that the data structure may play a role in limiting the ability to estimate some facet effects.

IV.5.2. Decision Studies

Before moving to describe the more complex models that I estimate in order to "adjust" for hidden facets, I describe the use of decision studies in GTheory. A GTheory analysis provides a variance decomposition of observed scores. The variance can be broadly broken down into True Scores, usually measured as only the teacher

[26] The specific condition here is that $\exists\ g\ st\ F_{\upsilon}\left(\sqrt{n}\left(g(\psi(\hat{\upsilon})) - g(\psi(\upsilon))\right)\right) \sim N(0,1)\ \forall \upsilon$ where g is monotonically increasing, $\psi$is the function of the parameters of (i.e. statistic being bootstrapped), $N(0,1)$ is a standard normal distribution, and υ is the vector of model parameters defining the distribution F (Efron & Tibshirani, 1994). If the model distribution truly fits the population, then confidence intervals will be more exact.

[27] This is not completely true because if all raters share the same bias, it will not be detectable without some external anchor. This is the problem of external rater error (Myford & Wolfe, 2009). It is also possible that a specific design is simply not well set up to distinguish rater-by-teacher bias from rater-by-day biases.

facet (i.e. $var(\upsilon_t)$), and Error, usually measured as all other facets. This can be used to generate an estimate of reliability (i.e. $var(TrueScore)/var(TotalScore)$). Importantly, a reliability can be estimated not only for the current data, but for alternative study designs (e.g. if teachers were observed on 5 days by 2 raters per day)[28]. Using bootstrapped samples, I can also estimate uncertainties for these reliabilities. One distinction often made in GTheory is whether teachers are compared to a fixed standard or to each other, called absolute and relative reliability, respectively. When comparing teachers to each other, under the assumption of fully-crossed raters and items, the rater and item main facets (i.e. $\upsilon_i,\ \upsilon_r$) do not contribute to measurement error. This is because the object of measurement is the teacher ranking, which remains unchanged if a rater or item becomes more or less lenient. However, observation instruments rarely have designs with fully crossed raters and changes to rater leniency will shift the relative rankings of teachers when the design is not fully crossed. Thus, the relative reliability is almost never appropriate for observation scores and I only use absolute reliabilities in this thesis. When displaying results from Decision Studies, I will generally graph the reliabilities across a range of days potentially observed and a range of raters potentially scoring each day, assuming 3 occasions per day for CLASS and PLATO and 1 occasion per day for FFT and all items scored.

IV.5.3. System Design Model (SD) The Base model described above forms the base for more complicated models that adjust for the three classes of hidden facets described earlier. The System Design Model (SD) described here adjusts for System Design Facets, including scoring mode (i.e. whether scoring was live or part of double scoring), rater drift

[28] This is possible because the variance of an average of two independent random variables can be easily calculated (i.e. $var((X+Y)/2) = (var(X) + var(Y))/4$ if $X \perp\!\!\!\perp Y$). Thus, if the variance of the rater facet is 0.1, then the error variance contributed by raters to a score averaged across two raters is 0.05 (i.e. $(0.1 + 0.1)/4 = 0.05$). This same type of analysis can be conducted across all facets after specifying the number of occasions, days, sections, items, and raters are being averaged over, allowing an estimate of the error variance and ultimately score reliability for a specific sampling design.

(i.e. date rater scored video), day of the week scored, date videotaped, and occasion order effects. This model replaces μ in in the Base model (equation Base) with:

$$\mu = \beta_{\text{Occ}\cdot\text{i}} + \beta_{Live} + \beta_{\text{Dbl}} + \beta_{\text{DtSc}} + \beta_{\text{DayWk}} + \beta_{\text{Month}} \qquad \text{(SD)}$$

where $\beta_{Occ\cdot i}$ captures occasion order by item effects (e.g. Positive Climate on segment 1), β_{Live} dummy codes whether scoring was done live, β_{Dbl} dummy codes whether scoring was part of the double scoring procedure, β_{DtSc} represents a linear trend for date scored, β_{DayWk} is dummy variables capturing day of the week videotaped (reference is Monday), β_{Month} is a linear trend for date observed.

The hidden facets controlled for in this model arise because the observation process must select specific days to observe using a specific scoring mode (i.e. live or by video) and raters score videos over time—all of which are determined by the observation protocol being implemented. This model forms one of the three models I contrast with the Base model in research question two. Generalizing observed scores should become more efficient after adjusting for these facets because sampling variation associated with these hidden facets is controlled for. Thus, teacher quality estimates should become more precise and the variance of the teacher facet (i.e. (υ_t)) will be reduced accordingly (because a source of sampling error included in this term is removed). Teacher quality, as defined in this model, represents the teacher's enacted teaching quality at a fixed occasion in time, using a given mode of scoring, and being scored at the same time[29]. That is, if teachers differ in quality based on when they were observed, the mode of the observation, or when raters scored their videos, these differences are removed from the estimate.

IV.5.4. Curriculum and Instruction Model (CI) The Curriculum and Instruction Model (CI) builds upon the System Design model by adding statistical adjustments for facets

[29] Note, however, adjustments for date scored and drift are only for average scores and assumed linear, which makes this definition an over-simplification.

related to the curriculum and instruction. These facets come from the PLATO log and include content domain taught and interaction structure. As with the System Design model, this model replaces μ in in the Base model (equation Base) with:

$$\begin{aligned}\mu = \beta_{\text{Occ·i}} + \beta_{Live} + \beta_{\text{Dbl}} + \beta_{\text{DtSc}} + \beta_{\text{DayWk}} + \beta_{\text{Month}} \\ +\beta_{\text{Read}} + \beta_{\text{Lit}} + \beta_{\text{Write}} + \beta_{\text{Grammar}} + \beta_{\text{Disc}} + \beta_{\text{Ind}} + \beta_{\text{Rec}}\end{aligned} \tag{CI}$$

where β_{Read}, β_{Lit}, β_{Write}, and $\beta_{Grammar}$ are dummy variables representing the four content domains: reading, literature, writing, and grammar, respectively; β_{Disc}, β_{Ind}, and β_{Rec} are dummy variables representing the three interaction structures: discussion, independent work, and recitation/lecture, respectively. These seven variables capture a range of content areas and instructional choices that most teachers likely use at some point during the year. However, they represent only a small subset of the important facets of Curriculum and Instruction that could be studied in research on teaching.

This model adds hidden facets related to the specific content and instruction occurring on the day observed to the SD model, helping to address research question two. To the extent that these facets act within-teachers, controlling for these facets should eliminate the sampling error associated with how often teachers are observed at each level of the facet. This should increase the precision with which teacher quality is estimated (and so reduce the variance at the teacher level), while the meaning of teacher quality shifts slightly. Teacher quality is now the teacher's capacity to engage in high quality instruction for reading, writing, literature, and grammar lessons while using discussions, independent work, and recitations on a given occasion in time, scoring mode, and when scored on the same day. That is, if teachers differ in teaching quality based on how often they are observed teaching across content domains, interaction structures, or SD facets, these differences are removed from score estimates. There may also be between-teacher aspects to these facets, which, depending on their cause, can lead this model to either increase or reduce bias in teacher quality estimates. Recall that even when this model increases bias by removing between-teacher differences in the frequency of engaging lessons at each level of the CI facets, the increase in precision may make this model appropriate because too few days of instruction are observed to accurately estimate how frequently teachers engage in specific types of instruction.

IV.5.5. School Organization Model (SO) The School Organization Model (SO) builds upon the Curriculum and Instruction model by adding statistical adjustments for facets related to the ways schools are organized. These facets reflect mostly between-teacher effects that reflect differences in grade, student composition, and school culture. As with the other models, this model replaces μ in in the Base model (equation Base) with:

$$\begin{aligned}\mu = {} & \beta_{\mathrm{Occ\cdot i}} + \beta_{Live} + \beta_{\mathrm{Dbl}} + \beta_{\mathrm{DtSc}} + \beta_{\mathrm{DayWk}} + \beta_{\mathrm{Month}} \\ & + \beta_{\mathrm{Read}} + \beta_{\mathrm{Lit}} + \beta_{\mathrm{Write}} + \beta_{\mathrm{Grammar}} + \beta_{\mathrm{Disc}} + \beta_{\mathrm{Ind}} + \beta_{\mathrm{Rec}} \\ & + \beta_{\mathrm{7th}} + \beta_{\mathrm{8th}} + \beta_{\mathrm{PrAch}} + \beta_{\mathrm{Demo}} + \beta_{\mathrm{Imp}}\end{aligned} \tag{SO}$$

where β_{7th} and β_{8th} are dummy variables capturing 7th and 8th grade (reference is 6th grade); β_{PrAch} is a linear effect for section average student prior achievement; β_{Demo} is a linear effect of a composite indicator of student demographics (discussed below); and β_{Imp} is a dummy variable capturing whether the prior achievement and student demographics were imputed (6% of classrooms). Imputations were done using a k-Nearest Neighbors algorithm using the *VIM* package in R (Kleiner, Talwalkar, Agarwal, Stoica, & Jordan, 2013). The composite indicator of student demographics is the first principal component of the section-level variables percent black, percent Hispanic, percent white, percent Asian, percent English language learner (ELL), and percent free-reduced price lunch (FRL). The first principal component explained 43% of the total variance and captures classrooms that are more black, Hispanic, ELL, and higher FRL while being less white and Asian.

This model adds hidden facets related to the teacher's context to the CI model. It is directly related to research question two. These facets should mostly act between teachers, affecting the extrapolation of scores across contexts. The extrapolation argument implicit in this model here is the co-construction argument (whereas teacher sorting is the implicit argument when not adjusting for these facets). Under the assumption of co-construction, correcting scores for student characteristics and grade taught is necessary to allow teacher scores to be compared across contexts. Teacher quality, as defined in this model, represents the teacher's ability to teach a specific type of classroom in a specific grade and to teach specific content domains using specific interactions structures at specific times. That is, differences in teacher quality associated with the students a teacher teaches, the grade at which they teach, or CI and SD facets are removed from estimates. Note that, under the teacher sorting assumption, adjusting for SO facets will introduce bias to teacher score estimates.

IV.6. Analyses

In this section, I detail the analyses that I will be conducting using the models just described. These analyses address research questions (RQ) 2 and 3 (whereas RQ 1 can be addressed using only the Base model). I first describe the model comparisons that address research question two, highlighting how comparisons across the four models just described can show the role of hidden facets in estimating teacher quality. I then discuss the third research question, focused on the validity of teacher score estimates across models and across levels of the hidden facets.

IV.6.1. Model Comparisons (RQ 2) There are three types of analyses that I will conduct to address RQ 2. These analyses compare the four models just presented (i.e. Base model, SD model, CI model, SO model) and estimate the impact of the hidden facet adjustments on observed teaching quality and estimates of teacher quality. The first set of analyses focus on the significance and size of the fixed effect estimates of the facets. The fixed effect estimates represent the impact of the facets on observed scores. They show how much scores might vary if, for example, a teacher is only observed at the beginning of the year compared to the end of the year. The size of the effect is difficult to interpret, though, because there is no meaningful metric. The natural metric of "scale point" says little about how much a teacher's score might be impacted by these effects. In order to create a meaningful metric, I convert the effects into an "effect size metric" using the standard deviation of the teacher scores from the Base model to scale effects (i.e.$\sqrt{var(v_t^{Base})}$). Note the use of a superscript to denote which model the parameter is from. This translates the facet effect into teacher standard deviation units (SD_T). For example, being scored live might move an observed score half of a teacher standard deviation (i.e. 0.5 SD_T), which would move a teacher from the 50th percentile of estimated teacher quality to the 69th percentile of estimated teacher quality. This also creates an arguably common unit across the models for different observation instruments. The reader will notice that the effect sizes can be quite large. This is a function of both the compressed range of observed scores and the relatively small percentage of variance that is attributable to teachers (i.e. high measurement error).

In order to address RQ 2a regarding instrument bias, I will compare estimated effects of hidden facets across models. When effects differ significantly across models (see Appendix F for a broader discussion of the statistical test used here), as I have argued, it is a sign of instrument bias, though determining which instrument is biased is impossible. In order to explore cases of instrument bias, I use item-specific GTheory models (presented in Appendix D) to analyze which specific items appear to be the source of the bias. If similar items across instruments show dissimilar effects of the hidden facet, then bias results from the relative emphasis that the different instruments place on specific aspects of teacher quality (i.e. construct under-representation or construct-irrelevant variance). If no such patterns exist, I will not be able to make any conclusions about the source of the bias.

As I've argued before, one of the main determinants of whether adjustments should be made for the effects of a hidden facet on observed teaching quality is whether the effect is between-teachers or within-teachers. This is RQ 2b. This is especially important for the

Curriculum and Instruction facets, which can affect scores both within-teachers between-days and between-teachers. I test for the level of the effect from hidden facets by dividing the hidden facet variable into three components: a within-teacher component, a between-teacher within-school component, and a between-school component. The within-teacher component is created by removing the teacher average score from the hidden facet variable. The between-teacher within-school component is created by removing the school average from the teacher average of the hidden facet variable. The between-school component is the school average of the variable. The within-teachers component is independent of teachers and schools and so can only act within-teachers. Similarly, the between-teachers within-school component is independent of schools and has a constant value within-teachers so can only act between-teachers, within-schools. The between-school component is constant within schools so can only act between schools. This centering trick has long been used by Hierarchical Linear Modeling approaches to explore the level at which variables act (Raudenbush & Bryk, 2001). This helps to clarify the nature and type of effects the hidden facets are having on observed scores. As I've argued before, adjusting for within-teacher effects should mostly increase the precision of estimates; between-teacher effects define where extrapolation across facets is necessary; and between-school effects both define where extrapolation is necessary and complicate extrapolation by conflating the hidden facet and broader school effects on teaching quality.

Turning to RQ 2c, I also conduct a number of analyses examining the change in teacher quality estimates across models. The size of the fixed effects does not directly show how much teacher quality estimates will change across models. That is because day-level facet effects on observed teaching quality scores are averaged across four days of instruction and effects from numerous facets. The net effect of facets varies in complex ways based on the distribution of facets across days and teachers. While the fixed effect estimate show the potential impact of facets on estimates of teacher quality, the difference in teacher quality estimates across models (i.e. comparing υ_t^{Base} and υ_t^{SD}) shows the actual impact on teacher quality estimates. The correlation of teacher quality estimates across models (i.e.$cor(\upsilon_t^{Base}, \upsilon_t^{SD})$) provides an overall estimate of how adjusting for hidden facets changes teacher scores. However, observation scores, in high stakes situations, can be used to make decisions about individual teachers (in combination with other data). Thus, knowing how much teacher quality estimates for individual teachers vary across models is important. This is often explored with classification consistency (Deng & Hambleton, 2013), which tests whether two models would classify teachers in the same way. However, classification consistency requires a threshold for comparison and there is no natural threshold that can be used in this case. Further, the use of a threshold becomes statistically complex when models adjust for facets. Thus, I will explore how teacher rankings change across models, which is akin to a threshold-less version of classification consistency. For example, a teacher may have an estimated score in the 50th percentile in the Base model and the 30th percentile in the System Design model, a difference of 20 percentile points. Looking across teachers, we might conclude that 5% of teachers have their scores shift 20 or more percentile points across models. This gives an estimate of how much individual teacher scores might be affected by adjusting for hidden facets.

Next, I look at the change in the variance of planned facets of measurement across the different models. As the adjustments for hidden facets are added to models, they will explain some of the variation across the different facets. This shifts the relative size of the true and error facet variances. A direct examination of the size of the shift in variance provides information about the improvements in precision gained by controlling for hidden facets. For example, assume the variance of the teacher facet is 20% smaller for the SD model as compared to the Base model. This would imply that 20% of the "teacher effect" from the

Base model is truly sampling variation stemming from the fact that teachers were observed at different times, using different observation modes, and raters scored videos over time. This means that the teacher quality estimate from the Base model is actually 20% sampling error.

The conclusion here assumes that the facets controlled for are all truly sources of measurement error and the shift in meaning of teacher quality, in this case from average provided teaching quality across a set time period to teaching quality on a specific occasion, observation mode, and scored on a given day. We can also explore changes in the size of the variance in the planned error facets, which provides information about how much the hidden facets included in the model explain the variation in observed teaching quality across the error facets. For example, if the SD model had 50% less rater error than the Base model, this would imply that half the rater error is attributable to scoring mode or rater drift (or the other SD hidden facets). This analysis provides another view into the importance of the hidden facets for observed teaching quality.

Additionally, the change in the variance associated with the different error facets across models will lead to different estimates for score reliability. Comparing the reliability across models tells how much the confidence we have in teacher scores changes after adjusting for the effect of hidden facets (RQ 2c). Importantly, the reliability in the adjusted model is reliability for a score with the same adjustments. This is important because, as discussed before, the meaning of the teacher quality estimate varies based on which hidden facets are adjusted for. Adjusting for hidden facets should decrease the teacher variance, removing sampling error from teacher score estimates, so the adjusted models will likely reduce reliability of scores. However, this is not necessarily the case. If adjustments explain significant amounts of the variance in error facets (e.g. the rater, item, or within-teacher section or day variance), the reliability may actually increase as more hidden facets are controlled for (i.e. the error variance may decrease).

The set of analyses described in this section provide evidence to address the second research question. They provide a comprehensive evaluation of the effects that hidden facets have on the measurement of teacher quality with observation scores. The results should help demonstrate when hidden facets have a meaningful effect on observed teaching quality and elucidate exactly what this effect is.

IV.6.2. Validity Analyses (RQ3) The analyses discussed up to this point emphasize changes to the reliability of scores and score estimates that occur after controlling for hidden facets. They have not yet dealt with the validity of scores, the third research question. In general, the validity of a score estimate is supported if it is related to similar measures of the same construct (i.e. concurrent validity). In this thesis, I use VA scores as the concurrent measure of teacher quality. That is, I argue that the validity of estimated teacher scores is stronger if it has a stronger relationship (such as a correlation) with VA scores. Further, if teacher score estimates are becoming *more valid* after adjusting for hidden facets (i.e. the scores are better capturing teacher quality and hence contain less error), the relationship between score estimates and VA scores should increase after statistically adjusting for more hidden facets. However, recall that VA score estimates and observational estimates of teacher quality both correlate with students' prior achievement, raising the concern of correlated measurement error. Thus, I test the relationship between estimates of observed teacher quality and VA scores after partialling out the effect of students' prior achievement and student demographics through simple OLS regression. VA scores (i.e. Y_{VA}) are the dependent variable while students' prior achievement, the demographic composite, and observation score are regressors (i.e. $Y_{VA} = \beta_{PrAch} + \beta_{Demo} + \beta_{v_t^{base}}$). If scores become more valid after adjusting for hidden facets, the $\beta_{v_t^{base}}$ term should increase across models (i.e. $\beta_{v_t^{base}} < \beta_{v_t^{SD}} < \beta_{v_t^{CI}} < \beta_{v_t^{SO}}$).

There is a second validity concern that arose in the theoretical framework and literature review. The relationship between observed teaching quality and teacher quality may vary across facets. This is a problem of differential validity across facets. Estimates of teacher quality may be more valid when created from observations of some facets than for other facets. For example, when observing small group instruction, there may be a strong relationship between estimated teaching quality and teacher quality while, when observing lectures, there may be a weak relationship between estimated teaching quality and teacher quality (i.e. $cor(\hat{\upsilon}_t, Y_{VA})|smallgroup > cor(\hat{\upsilon}_t, Y_{VA})|lecture$). This can occur if the observation instrument measures interpersonal interactions, but, for lectures, teacher quality is more related to the organization of content than interpersonal interactions. That is, the observation instrument measures some aspect of instruction that is somewhat tangential to teaching quality on some levels of a facet. I explore the possibility of differential validity for the Curriculum and Instruction facets and the School Organization facets by testing whether the estimated teaching quality score-VA score relationship is affected by how often teachers were observed on a given hidden facet. For example, if observation scores of writing lessons are more valid than those of non-writing lessons, the validity of observation scores should be higher when more writing lessons are observed (i.e. $\beta_{\upsilon_t^{base}} \Big| small\ group\ always\ observed > \beta_{\upsilon_t^{base}} \Big| small\ group\ never\ observed$ where $\beta_{\upsilon_t^{base}}$ is as defined in the validity equation in the last paragraph). This can be tested by interacting observation scores with a variable capturing how many days of writing were observed for the given teacher in a model predicting VA scores. The use of VA scores as a validation measure is common practice in research on classroom observation instruments, but likely has little power given the weak relationship between observation score estimates and VA scores and the distance of VA scores from classroom instruction. The difference in validity across models and across hidden facets therefore would have to be very large to have any power to detect the kinds of effects just noted.

IV.7. Summary

This chapter reviewed the data sources, statistical models, and analytic approaches used in this thesis. The UTQ project provides a rich source of data to explore the measurement properties of classroom observation scores as measures of teacher quality. GTheory provides the statistical framework to explore the properties of classroom observation scores, allowing the separation of true teacher effects from multiple sources of planned error. Further, GTheory is easily expanded to account for the situated nature of teaching and the effect of hidden facets on the reliability, bias, and validity of estimates of teacher quality.

I described three approaches to explore the effect of hidden facets on observed teaching quality in this chapter. First, I examine the size of the fixed effect estimate, scaled to an effect size metric, which shows the effect of the hidden facet on observed teaching quality. Further, differences in effect sizes across instruments indicate instrument bias. Second, the correlation of teacher score estimates and the shift in ranks of those estimates demonstrate how much correcting for the effects of hidden facets actually changes estimated teacher scores. Third, the change in score reliability across models shows how the amount of error in score estimates shifts after controlling for hidden facets. I further described how I will divide hidden facets into independent components across levels of nesting (i.e. within-teachers, between-teachers, between-schools) to explore the level of nesting at which hidden facets affect observed teaching quality. Last, I examine the differential validity of scores across facets and across different degrees of adjustments for hidden facets.

These analyses provide a comprehensive view of how adjusting for hidden facets affects observed teaching quality and teacher quality estimates. It is not clear, though, what degree of adjusting is ideal. In fact, the very notion of an ideal set of facets to adjust for is probably overly-simplistic. The types of adjustments one decides to make depends on definitions of teacher quality, what facets one believes are drivers of teacher quality, and whether there are any practical effects of making adjustments.

Chapter V. Results

In this chapter, I report the results from the analyses just described. In the first section, I focus on the relative size of the error facets included in the base GTheory model (Research Question [RQ] 1). This provides a broad overview of how the observation instruments are functioning as tools of measurement across the many sources of planned error inherent to the measurement protocol. I then turn to reporting the fixed effect estimates of the hidden facets in the System Design (SD) model, Curriculum and Instruction (CI) model, and School Organization (SO) model, paying special attention to the differential effects of hidden facets on observed teaching quality across instruments. These results address RQ 2 broadly. In the same section, I also explore whether the effects of hidden facets on observed teaching quality are within-teachers or between-teachers (RQ 2b), which has important implications for how this might affect estimates of teacher quality. In the third section, I explore the impact that hidden facets have on estimates of teacher quality, including the reliability of these estimates (RQ 2c). I then turn to the problem of validity by looking at how the teacher quality estimates from different statistical models are correlated to UTQ value-added scores (RQ 3).

V.1. Results from the Base Model

In this section, I review the results of the Base model. This model estimates the relative contributions of teacher quality (i.e. the "true" score) and error facets to variance in observed teaching quality scores. The goal is to evaluate RQ 1, which asks about the relative size of the contributions of teacher quality and the error facets to observed teaching quality. Developing an understanding of the relative importance of the different error facets is an important first step in understanding how observed teaching quality varies across contexts of measurement.

Table 5.1 shows the size of the variance of the planned facets of measurement from the Base models. The Base model was estimated separately for each of the three instruments, and the results for each instrument are presented in two columns. The left column under each instrument shows the absolute size of the estimated variance components while the right column presents the same data as a percentage of the total variance attributable to each facet. Both columns contain 95% bootstrapped confidence intervals in parentheses below the estimates. The column presenting percentages of variance explained by facets is generally more useful because it scales the instrument-specific variance components to a common and meaningful scale.

I will start by discussing the estimated variance of teacher quality (i.e. the teacher facet or "true" score in each model). I will then turn to a discussion of the day and occasion error facets, also discussing here the item error facets related to days and occasions. I then review the item main effect facet, the results of which are displayed (separately) in Table 5.2. Last, I will discuss the rater error facets.

Table 5.1: Random Effect Variance Components from Base GTheory Model

	CLASS		FFT		PLATO	
Facet	Value	Percent	Value	Percent	Value	Percent
Teacher ($var(v_t)$)	0.076 (0.054-0.098)	7% (5-8.9)	0.029 (0.021-0.038)	10.7% (7.9-13.7)	0.012 (0.007-0.015)	2.8% (1.8-3.7)
Day ($var(v_{d:s:t})$)	0.013 (0-0.035)	1.2% (0-3.2)	0.008 (0.001-0.016)	3% (0.4-5.9)	0.003 (0-0.008)	0.8% (0-1.8)
Occasion ($var(v_{o:d:s:t})$)	0.053 (0.048-0.059)	4.9% (4.3-5.5)			0.017 (0.015-0.019)	4.1% (3.6-4.6)
Rater ($var(v_r)$)	0.04 (0-0.102)	3.7% (0-8.9)	0.011 (0.003-0.024)	4.2% (1-8.5)	0.002 (0-0.009)	0.5% (0-2.2)
Rater-by-Teacher ($var(v_{rt})$)	0 (0-0.037)	0% (0-3.3)	0.005 (0-0.019)	1.8% (0-6.6)	0 (0-0.005)	0.1% (0-1.2)
Rater-by-Day ($var(v_{r(d:s:t)})$)	0.141 (0.099-0.156)	13% (9.1-14.6)	0.044 (0.029-0.054)	16% (10.5-19.6)	0.02 (0.014-0.024)	4.8% (3.5-5.7)
Item-by-Rater ($var(v_{ir})$)	0.225 (0.17-0.294)	20.7% (16.4-25.7)	0.011 (0.008-0.015)	4% (2.9-5.2)	0.022 (0.014-0.03)	5.3% (3.5-7.1)
Item-by-Teacher ($var(v_{it})$)	0.029 (0.024-0.034)	2.7% (2.2-3.2)	0.008 (0.006-0.01)	2.9% (2.1-3.7)	0.012 (0.009-0.015)	2.9% (2.3-3.6)
Item-by-Day ($var(v_{i(d:s:t)})$)	0.128 (0.12-0.135)	11.8% (10.7-12.8)	0.017 (0.012-0.021)	6% (4.5-7.7)	0.067 (0.062-0.07)	16.2% (15.2-17.1)
Item-by-Occasion ($var(v_{i(o:d:s:t)})$)	0 (0-0.008)	0% (0-0.7)			0.012 (0.007-0.017)	2.9% (1.7-4.1)
Residual ($var(\epsilon_{ir(o:d:s:t)})$)	0.381 (0.372-0.386)	35.1% (32.3-37.4)	0.14 (0.135-0.145)	51.4% (48.1-54.3)	0.246 (0.241-0.251)	59.6% (57.7-61.4)

Note. Separate regressions were run for each instrument. For each regression model, the value column shows the estimated variance for the given facet and given model and the percent column shows the percentage of the total variance contributed by the given facet. * p<0.05; ** p<0.01; *** p<0.001.

V.1.1. Teacher Facet (Teacher Quality) Table 5.1 shows that the percentage of variance attributable to the teacher quality (i.e. the teacher facet: v_t) was 7% for CLASS, 10.7% for FFT, and 2.8% for PLATO. Thus, across all three instruments, 11% or less of the variance in observed scores was attributable to teachers—the object of measurement in classroom observation research. The reader should note that this is much lower than what has been found in past research. For example, the MET study (discussed earlier in this thesis) found that the percentage of variance due to the teacher facet, was about 30% for the same instruments examined here (c.f. Kane et al., 2012). The difference between these MET results and the ones presented in Table 5.1 was mainly due to differences in the GTheory statistical model estimated here versus statistical model in the MET study rather than inherent properties of the UTQ data set. As discussed earlier, the GTheory statistical models I estimated included items as facets of measurement, which makes comparisons with the item-averaged statistical models used in the MET study inappropriate. If we estimate the same statistical models employed by MET researchers using UTQ data, the results were more similar to results from that study—24% variance in observed scores was due to teacher quality for CLASS, 27% for FFT, and 22% for PLATO.

The results in Table 5.1 show that the percentage of variance due to teachers differs significantly across instruments. FFT had the highest percentage of variance in scores due to teachers (10.7%); by contrast, the percentage of score variance due to teachers was about a third as much for PLATO (2.8%) and somewhat more than half as much as for CLASS (7%). This shows that compared to CLASS and PLATO, the variance in observed scores on FFT were more the result of teacher quality (v_t in Equation 1) than to the effects of the rater, item, day, or occasion facets in the model (alone and in combination). As a result, score estimates from FFT should be more reliable (but not necessarily more valid) than scores from CLASS or PLATO.

Note also that the percentage of variance in observed scores due to teachers was estimated with minimal "absolute error", but considerable "relative error". For example, Table 5.1 shows that the confidence intervals for the percentage of variance due to teachers were only a few percentage points wide, which shows a small absolute error (i.e. the confidence interval spans only a 2-5 percentage points). However, the uncertainty in the estimate was large relative to the size of the point estimate (i.e. relative error). For example, the variance point estimates shown in Table 5.1 can shift up or down by about 33% and still remain within the 95% confidence interval. This is an important challenge to my subsequent efforts to understand how *hidden* facets affect the measurement properties of observation instruments. As I have argued, one of the likely main effects of hidden facets in GTheory analyses is to inflate estimates of the variance of the teacher facet ($\widehat{var}(v_t)$). As I turn to making comparisons between models to explore this effect, one question will be whether the teacher-level variance component was reduced as adjustments were made for more hidden facets. Given the confidence intervals for variance components shown in Table 5.1, hidden facets would have to explain over 1/3 of the variance between teachers in the Base model for me to conclude that a statistically significant change has occurred in the estimated variance of teacher quality. In fact, as can be seen in Table 5.1, this is a challenge not only for the teacher variance component, but for all other variance components in the model as well.

V.1.2. Day and Occasion Error Facets To this point, I have been discussing the variance of the "true" score component in my GTheory statistical model ($\hat{var}(v_t)$). I now turn to the percentage of variance in observed scores due to specific error facets, beginning with the day and occasion error facets. The day facet produces variation in observed scores within-teachers between-days, while the occasion facet produces variation in observed scores within days. Table 5.1 shows that all three of the classroom observation instruments under study had low day variance, with the day facet accounting for only 1.2% of the total variance in observed scores for CLASS, 3% for FFT, and 0.8% for PLATO. This is surprising given that past work has found larger effects for the day facet (e.g. Kane et al., 2012). Just as for the teacher facet, however, part of the discrepancy in my findings versus those of other studies stems from the model estimated in this thesis, which led to lower estimates of variance components generally. However, even after estimating a MET-like model on UTQ data (results *not* shown here), the variation in observed teaching quality across days in UTQ data was below that found in the MET data. This is unfortunate in the context of this dissertation because many of the hidden facets of interest to this dissertation (like curriculum and instruction) varied across days. Because there was low within-teacher variation in scores by day, there was less variation to be explained, making the detection of "hidden" facet effects more difficult when these were included in my statistical model.

Table 5.1 shows that the variance due to occasions (nested within days) was larger than the variance due to days, with the occasion facet accounting for 4.9% of the total variance in CLASS and 4.1% in PLATO. Note again that FFT was scored on 30 minute occasions and that there were not enough days scored on multiple occasions to estimate an occasion effect. As a result, no estimate of occasion variance was provided for FFT in Table 5.1. Overall, the findings on the occasion variance seem to suggest that occasions were more important than days in accounting for observed score variance. However, it is important to consider the item-by-day and item-by-occasion effects before drawing this conclusion. The percentages of variance in observed scores due to item-by-day effects were large for all instruments: about 11.8% for CLASS, 6% for FFT, and 16.2% for PLATO. In contrast, the percentage of observed score variance due to item-by-occasion effects was much smaller: 0% for CLASS and 2.9% for PLATO. Thus, when including item effects, days were a more important source of variation in observed teaching quality than occasions. This shows that items were very important for understanding the variation of teaching quality across days, but less important for understanding how occasions deviate from day scores.

Using the data in Table 5.1 we can quantify the relative importance of average scores across items (e.g. the day facet) and deviations from this average due to specific items (e.g. the item-by-day facet) at each level (i.e. occasion, day, teacher) of the statistical model. The sum of the variance due to the day facet and the item-by-day facet represents the total variance in observed scores across days (net of any rater error effects). As the day facet gets relatively larger and starts to explain all of the variance in scores across days (i.e. $var(v_{d:s:t}) >> var(v_{i(d:s:t)})$), items do not vary independently, but only vary with changes to the day mean score (i.e. the day facet). That is, the variance across days becomes unidimensional. This suggests using the percentage of the variance across days (net of rater error effects) that is due to the day facet (i.e. $var(v_{d:s:t})/\left[var(v_{d:s:t}) + var\left(v_{i(d:s:t)}\right)\right]$) as a rough measure of uni-dimensionality (see "percentage of total variance" in Hattie, 1985). Applying this measure to data from Table 5.1, I find 9% of the variance in observed CLASS scores across days was due to the day facet; while the equivalent percentages for FFT and PLATO were 32% and 4%, respectively. Thus, there was a great deal of multi-dimensionality across days; that is, deviations of specific items from the average day score was instrumental to understanding the variation of observed scores across days, implying days vary in instructional quality across specific dimensions more than they do overall.

Conducting the same analysis focusing on the occasion facets, we can see that *all* of the variance in observed CLASS scores across occasions was due to the occasion facet and 59% of the variance in observed PLATO scores across occasions was due to the occasion facet. These percentages were much higher than for days, which show more unidimensionality at the occasion-level. That is, item-specific deviations from the average score were much less important for understanding the variance of observed scores across occasions than across days. This tells us that day deviations from teacher quality occurred mainly on specific items (i.e. a day was stronger or weaker than expected on specific items rather than as a whole) while occasion deviations from day scores occurred equally across all items (i.e. an occasion was stronger or weaker than expected equally across all items).

We can conduct this same analysis at the teacher-level (i.e. examine what percentage of the teacher variance was due to the teacher facet). Using the data from Table 5.1, I find that 70% of variance in observed CLASS scores across teachers was due to the teacher facet while the corresponding percentage was 80% for FFT and 50% for PLATO. Thus, like at the occasion-level, the variance of observed scores at the teacher-level was mostly due to differences in teacher means, rather than item-specific deviations from the mean (i.e. scores were more unidimensional at the teacher-level). There were, however, differences across the instruments; item facets contributed more to PLATO scores across all levels of the model (i.e. occasions, days, and teachers), as compared to CLASS and FFT. The variation in FFT scores was least effected by items at all levels (i.e. occasions, days, teachers).

V.1.3. Main Item Error Facet I have just discussed how item error facets interact with occasion, day, and teacher facets to produce observed score variance. I turn now to examining item main effects. Note that in my Base statistical model, items were modeled as fixed effects and so are not included as a facet in the variance components results shown in Table 5.1. Table 5.2 shows the item main effects with one pair of columns for each instrument. Estimates of item means with standard errors from the Base model are displayed for each item. As Table 5.2 shows, the item means spanned a wide range of the scale for each instrument. Moreover, if item means were treated as a random effect in my statistical model, they would dwarf the variance accounted for by any other facet, except the residual[30]. As has been noted in past work (e.g. Kane et al., 2012) and as shown in Table 5.2, teachers tended to receive higher scores on items measuring the classroom management and classroom culture dimensions of teaching quality and tended to receive lower scores on items measuring various instructional dimensions of teaching quality.

[30] As discussed before, by including item fixed effects in my Base model, I have effectively reduced the overall amount of variance to be explained by the random effects in my model. My treatment of item effects as fixed means that I assume any researcher building on my results to make a decision study will use all items in his or her observation protocol. Moreover, although I include item fixed effects in my Base model, the actual teacher quality score I get from my Base model (i.e., the specific random effect for a given teacher's estimated from the model) would correlate 1.0 with an estimate of the same model without these fixed effects. The benefits of including item fixed effects in the BASE model is that they are indicator of “item” difficulty. That is, item means show which items teachers tend to score high on and which they score lower on.

Table 5.2: Average Item Scores from Base GTheory Model

CLASS		FFT		PLATO	
Item (β_i)	Mean (SE)	Item (β_i)	Mean (SE)	Item (β_i)	Mean (SE)
Positive Climate	4.57 (0.15)***	Respect and Rapport	2.82 (0.05)***	Purpose	2.88 (0.07)***
Negative Climate	6.71 (0.15)***	Culture for Learning	2.35 (0.05)***	Intellectual Challenge	2.12 (0.07)***
Regard for Adolescent Perspectives	3.12 (0.15)***	Classroom Procedure	2.49 (0.05)***	Representation of Content	2.42 (0.07)***
Teacher Sensitivity	4.04 (0.15)***	Student Behavior	2.77 (0.05)***	Connections to Prior Knowledge	1.51 (0.07)***
Behavior Management	5.96 (0.15)***	Physical Space	2.34 (0.05)***	Connections to Personal Experience	1.31 (0.07)***
Productivity	5.73 (0.15)***	Communicating with Students	2.64 (0.05)***	Explicit Strategy Instruction	1.17 (0.07)***
Instructional Learning Formats	3.71 (0.15)***	Knowledge of Content and Pedagogy	2.24 (0.05)***	Modeling	1.24 (0.07)***
Content Understanding	3.26 (0.15)***	Questioning and Discussion Techniques	1.97 (0.05)***	Guided Practice	2.42 (0.07)***
Analysis and Problem Solving	2.42 (0.15)***	Engaging Students	2.27 (0.05)***	Classroom Discourse	2.07 (0.07)***
Quality of Feedback	3.36 (0.15)***	Using Assessment	2.04 (0.05)***	Text Based Instruction	1.93 (0.07)***
Student Engagement	5.02 (0.15)***	Flexibility and Responsiveness	2.15 (0.05)***	Accommodations for Language Learning	1.37 (0.07)***
				Behavior Management	3.91 (0.07)***
				Time Management	3.77 (0.07)***

V.1.4. Rater Error Facets Having considered teacher, day, occasion, and item facets, I turn now to facets of measurement involving raters. Table 5.1 shows that there are four error facets related to raters, plus the residual which captures rater error (since the residual includes rater-by-occasion, rater-by-item-by-teacher, rater-by-item-by-day, and rater-by-item-by-occasion effects). The rater facet in my statistical model captures variation in scores due to some raters being consistently more harsh or lenient than other raters in their scoring. The rater-by-item error facet captures an item-specific version of this same error (and if large, shows that a rater's leniency is not consistent across items). The rater-by-teacher and rater-by-day error facets capture idiosyncratic rater reactions to specific teachers and specific days, respectively[31]. Thus, each of the error facets captures a different type of rater error.

Table 5.1 shows wide variation in how much each type of rater error contributed to the total variance in observed scores across the three classroom observation instruments under study, although rater-by-item and rater-by-day error facets were always the largest error component, no matter the instrument. Looking at Table 5.1, it can be seen that the percentage of variance in observed teaching quality explained by the rater-by-item error facet is 20.7% for CLASS, 4% for FFT, and 4.8% for PLATO. The noticeably larger variance explained by the rater-by-item facet for CLASS versus the other instruments may be due to differences in the structure of CLASS instrument itself or to the rater training for CLASS, though because the same raters scored all three instruments in the UTQ study it cannot be due to the raters themselves. For both CLASS and PLATO, the rater-by-item error facet was the largest rater-related effect in the data (not including the residual), easily dwarfing the variance in observed scores due to the rater main effect only. This implies that on CLASS and PLATO, raters were not so much harsh or lenient in general but rather were relatively harsh or lenient in their scoring of specific items. For FFT, where item facets were generally small, the rater-by-item error facet was almost as large as the rater error facet. Thus, for all of the instruments under study, raters appeared to be more or less lenient (compared to other raters) on an item-by-item basis.

Table 5.1 also allows us to examine the importance of the rater-by-day effect, which (as the table shows) is the next most important source of rater error. The percentage of variance in observed teaching quality explained by the rater-by-day error facet is 13% for CLASS, 16% for FFT, and 4.8% for PLATO. This error facet is noticeably smaller for

[31] Recall that the rater-by-section error was found to be near zero so these error facets are independent of errors stemming from the specific classroom.

PLATO than it is for CLASS and FFT. Looking at Table 5.1, we can also compare the relative size of the rater-by-day facet and the day facet. Doing so shows that the rater-by-day facet was much larger than the day facet, implying that any two raters disagreed over the correct score on a given day more than any two days "disagreed" about a teacher's teaching quality[32]. No matter the causes of the rater-by-day error in the data, this indicates a very high-level of rater error in the estimation of teaching quality on a given day. This again raised a challenge for my thesis, especially for hidden facets that operated at the day-level. In UTQ data, most days were scored by a single rater, and as a result, day-level score estimates were confounded with rater effects, which affects my exploration of day-level hidden facets[33]. The UTQ data set, then, may be limited in its ability to explore the impact of (within-teacher) day-level hidden facets since scores for days were not well-estimated.

In contrast to the rater facets discussed so far, the rater-by-teacher facet was indistinguishable from zero (see Table 5.1). This finding helps alleviate common concerns that raters were biased against specific types of teachers. Since any such bias would show up in the rater-by-teacher facet (unless all raters were biased in the same way[34]), this concern

[32] One possible way of exploring this kind of error in UTQ data is to look at the number of notes that raters submitted for each occasion and day. UTQ asked raters to submit notes that recorded evidence they used to assign scores (although it is not clear how well this policy was implemented since there was a wide variation in how often raters submitted notes containing scoring evidence). I conducted an exploratory analysis examining variation in the number of score notes submitted by raters. After adjusting for rater main effects, the largest source of variation in scoring notes submitted was the rater-by-day facet (except for PLATO where the rater-by-occasion facet was slightly larger). Further, for CLASS and PLATO, days accompanied by more scoring notes had higher scores while days accompanied by more scoring notes had lower scores on FFT. Thus, the data indicate that raters apparently noticed different amounts of scoring evidence when observing the same day of instruction and difference in how much scoring evidence was reported is associated with differences in scores. This suggests that the ways in which raters confronted and processed evidence on a given day could be an important explanation of rater-by-day error.

[33] Rater-by-day effects and the assignment of raters to days might also explain the low section-level variances estimated in my initial models. Without being able to stably estimate day-level deviations from teacher scores, the model may be unable to estimate section average scores, especially with only two days per section. A true exploration of the section facet, then, might have to wait until a more robust data set is created that increases both the number of days scored for each teacher and the number of raters scoring any given day.

[34] In a set of analyses outside the scope of this dissertation, I show about half of the rater error in UTQ is the result of all raters being biased in the same direction (relative to master scores on calibration data). Thus, the rater error in analyses discussed in this thesis captures only half of the total rater error.

does not seem warranted in the UTQ data, perhaps because UTQ employed professional raters with no knowledge of the teachers they were rating[35].

In summary, this section discussed the effect that teacher quality and error facets had on observed teaching quality as estimated from the base GTheory model applied to UTQ data. Exploring the relative size and importance of the random effects from the GTheory model highlighted many of the ways that observed teaching quality varied over the measurement facets included in the analysis, providing useful information about errors in the measurement of teacher quality. There were differences across instruments in the importance of different facets, including differences in the amount of variance in observed teaching quality that was due to differences in teacher quality (i.e. size of teacher facet), the importance of items in understanding observed teaching quality, and the main types of rater error. On CLASS, the item-by-rater, item-by-day, and residual facets explained the largest percentage of variance in observed teaching quality while the teacher facet (i.e. teacher quality) explained a moderate amount of the variance in observed teaching quality. On FFT, the rater-by-day, teacher, and residual facets explained the largest percentage of variance in observed teaching quality. On PLATO, the rater-by-day, item-by-day, and residual facets explained the largest percentage of variance in observed teaching quality.

[35] Once again, it is worth noting that the UTQ data structure does not provide a strong basis from which to evaluate the rater-by-teacher facet and in fact may lead this facet to be under-estimated. Ideally, the rater-by-teacher facet is estimated when two different raters score all observed days for a given teacher. This allows each rater to generate a complete view of the teacher from which comparisons across raters can be made. In UTQ, raters very rarely scored more than two out of four days from a teacher, conflating the rater-by-day and rater-by-teacher errors. Surprisingly, the uncertainty in the rater-by-teacher variance components is not correspondingly large. It is interesting to note, however, that across individual bootstrapped samples, the correlation in the variance estimates for the rater-by-teacher and rater-by-day effects is quite large (near -0.8), confirming my suspicion that the data structure leads to poor separation of these effects. Simulation work is necessary to explore the effect of data structure on limiting how error facets can be estimated.

V.2. Impact of Hidden Facets on Observed Teaching Quality

In this section, I turn from investigating variance in observed scores due to planned features of the observation protocol to what I previously called “hidden” facets of measurement. Recall that I discussed three general “classes” of facets: System Design (SD) facets, Curriculum and Instruction (CI) facets, and School Organization (SO) facets. In what follows, I explore the impact of these facets on observed teaching quality ($X_{ir(o:d:s:t)}$) in a set of nested statistical models that progressively adjust for these facets, beginning with a GTheory model that adds to the Base model the effects of the SD facets, moving next to the incorporation of CI facets, and concluding with the incorporation of SO facets. The goal of these nested models is to address RQ 2 by estimating the extent to which observed teaching quality changes across levels of hidden facets. Specifically, I look across the three instruments to examine evidence of instrument bias across the hidden facets (RQ2a); I address the question of whether hidden facets act within-teachers, between-teachers, or between-schools (RQ2b); and I show how adjusting for hidden facets impacts estimates of the variance of teacher quality ($var(v_t)$) and the variance of the planned error facets.

Two challenges arose in comparing results across these nested models: determining whether facet effects on observed scores are meaningfully large (i.e. understanding effect sizes) and comparing facet effects across the three instruments. To address these challenges, I reported facet effects in the tables below in the metric of teacher quality standard deviations (i.e. $\sqrt{var(v_t)}$, which I denote below as SD$_T$). Standardizing on teacher quality standard deviations allowed me to interpret facet effect sizes in terms of how much entry of a given facet into my GTheory statistical model would move a teacher across the distribution of teacher quality. This, in turn, arguably created a common, meaningful metric across instruments[36].

V.2.1. System Design Model (SD) I begin the analyses of hidden facets by exploring the effect of the System Design (SD) facets on observed scores. This involves adding variables characterizing the SD facets to the Base model discussed earlier. In what follows, I call this new model the SD model. By demonstrating that the SD facets affected observed teaching quality (i.e. $X_{ir(o:d:s:t)}$), I show that the SD facets I consider are, in fact, hidden facets (recall that hidden facets must affect observed scores and capture a characteristics we wish to generalize across). I also look for evidence of instrument bias to address RQ 2a.

My findings show that a number of the System Design facets were systematically related to observed teaching quality (i.e. $X_{ir(o:d:s:t)}$) as shown in Table 5.3. Table 5.3 shows the results of three separate SD models (one for each instrument). Each cell contains the estimated effect of a hidden facet on observed teaching quality with the standard error of that effect. The top row of Table 5.3 shows that the effect of a dummy coded variable representing whether a rater scored a day live, that is, whether the rater was in the classroom or using a pre-recorded video (where live scoring = 1, and video scoring = 0). The table shows that live scoring had a statistically significant effect on observed scores only for the FFT instrument. For FFT, days scored live received scores 0.56 SD_T higher on FFT than days scored from video. This implies that a teacher at the 50th percentile of estimated teacher quality would be estimated to be at the 71th percentile of estimated teacher quality if they were only scored live. CLASS scores were not higher when scoring was live (β_{Live} =0.33 SD_T; p=0.11). In contrast to the two instruments just mentioned, PLATO scores were 0.44

[36] I present the same tables, but this time in the typical scale point metric, in Appendix B.

SD_T lower when scoring was live, an effect that was marginally significant (p=0.07). Considering FFT had a significant positive effect and PLATO had a marginally significant negative effect, the impact of live scoring on PLATO and FFT are inconsistent, suggesting instrument bias, as discussed earlier. Appendix F discusses the process of identifying this bias in more detail. Here, I simply state that the effect of live scoring on PLATO scores was significantly lower than the effect on FFT scores (p<0.001) and CLASS scores (p=0.008), demonstrating instrument bias. This indicates bias because live scoring cannot simultaneously lead to an increase and a decrease in true observed teaching quality (i.e. X_{true})[37]. Note that this bias is likely due to construct under-representation or construct-irrelevant variance[38]. That is, the biased instrument is not capturing some important aspects of true teaching quality (construct under-representation) or is measuring some factor that is independent of true teaching quality (construct-irrelevant variance) when scoring is live.

Table 5.3: Fixed Effects for the System Design (SD) Model across the three Instruments in the Teacher SD Metric

Names	CLASS	FFT	PLATO
Scored Live (β_{Live})	0.33 (0.20)	0.56 (0.21)**	-0.44 (0.25)
Double Scored (β_{Dbl})	-0.12 (0.13)	0.05 (0.13)	-0.24 (0.14)
Date Scored (m) (β_{DtSc})	-0.06 (0.01)***	-0.04 (0.01)**	-0.08 (0.02)***
Day of the Week (β_{DayWk})			
Tuesday	-0.08 (0.15)	-0.01 (0.14)	-0.16 (0.18)
Wednesday	0.31 (0.16)*	0.29 (0.15)	0.18 (0.19)
Thursday	-0.02 (0.15)	0.15 (0.15)	-0.20 (0.18)
Friday	-0.24 (0.18)	-0.15 (0.18)	-0.02 (0.23)
Observation Month (β_{Month})	-0.12 (0.02)***	-0.11 (0.02)***	-0.11 (0.03)***

Note. Each column shows the results of a separate model for the indicated instrument. Date Scored is scaled so a 1 point difference is one month; Monday is the reference group for the Days of the Week. * p<0.05; ** p<0.01; *** p<0.001.

[37] It is somewhat strange to think of the process of live scoring as affecting true observed teaching quality. However, if, for example, low-achieving students routinely sit in the back of the classroom, where they are observable only when live scoring and not on video, then the process of scoring classrooms live could have a real impact on the true value of observed teaching quality by allowing a unique aspect of the classroom to be visible. Alternatively, the presence of a rater scoring the classroom could affect what happens in the classroom more than the presence of a video camera. Either of these could differentially affect different dimensions of observed teaching quality and so have a different effect across instruments.

[38] In an analysis not shown here but reported in Appendix D, I ran the SD model separately for each item in each instrument. These item-specific models allowed me to explore whether the inconsistency in effects of the type of scoring across PLATO and FFT was restricted to certain items on these instruments. However, no clear patterns emerged in the data.

The second row of Table 5.3 shows the effects of whether or not a score was part of the double scoring process on observed scores. As the table shows, this facet had no statistically significant effect on observed scores[39]. The third row of Table 5.3 (i.e. the Date Scored row) shows the effect of when scoring was completed on observed teaching quality. This variable is included in the SD model to capture a phenomenon called "rater drift" (where any linear trend in scores across dates indicates a trend in scoring that is, by design, independent of other explanations for this trend). As Table 5.3 shows, all three instruments showed negative rater drift with raters becoming harsher (i.e. giving lower scores) over time. Further, the estimated size of the rater drift was consistent across instruments. While the rater drift effect is small, scoring persisted over a two year period and the effect shown gives the difference in assigned scores across adjacent months. Thus, the difference between scores given at the start and end of the scoring process due solely to the effect of rater drift is about 1.4-1.9 SD_T, which is quite large. The finding of negative rater drift matches the results of Casabianca and colleagues (2015), who showed a similar effect in the UTQ data, though they modeled a complex drift trend that arguably over-fits the data.

Turning from SD facets related to how and when observations were scored, I next examine the effects of when the instruction being observed took place. I begin by looking at results in Table 5.3 showing day of the week effects on observed scores. Four effects for days of the week were estimated (with Monday as the reference day). As Table 5.3 shows, scores on CLASS and FFT were higher on Wednesdays (~0.30 SD_T) compared to other days of the week, though this effect was only marginally significant for FFT. This Wednesday effect

[39] The effect of double scoring is marginally significant for PLATO. It becomes significant in later models. There is no reason for there to be a significant effect of double scored videos. Double scored videos were randomly selected, scored in a random order (albeit later on average than the original scores), and scored by a randomly selected rater. It is possible that the correction for rater drift is not adequate, but adding a more complex drift term did not affect the significance of the double scoring effect. Looking at the item-level data, the effect is driven by only Modeling in the SD model while Purpose, Intellectual Challenge, and Representation of Content become significant in the CI model.

would move an average teacher from the 50th percentile of estimated teaching quality to the 62th percentile. In contrast to CLASS and FFT, there were no day of the week effects for the PLATO instrument. Here again, then, there was some suggestion that instrument bias may exist. However, the effect of Wednesdays on CLASS scores was not significantly different than the effect on PLATO so instrument bias cannot be confirmed (c.f. Appendix F).

Table 5.3 also shows the effect of the month that an observation took place (denoted as Observation Month in the table). All three instruments show a decrease in observed teaching quality scores over the course of the year, suggesting actual teaching quality decreases across the school year. Table 5.3 shows that CLASS, FFT, and PLATO scores decrease ~0.11 SD_T for each month of the school year. This negative effect on observed teaching quality has been found before with UTQ data (Casabianca et al., 2015). Over the 8 month school year, the trend in scoring predicts that scores will decrease by 0.99 SD_T on CLASS, FFT, and PLATO, which can substantially affect a teacher's score.

A final step in the analysis of SD facets is reported in Tables 5.4 and 5.5. Here, attention turns from a consideration of time of year and day of week effects on observed scores to the effects of occasions of measurement within days on observed scores. In these tables, I am going to report on the extent to which observed scores on CLASS and PLATO are affected by the time ordering of observation segments within days. The reader will recall that raters using these instruments recorded their scores at 15-minute intervals[40]. The question for the analysis is whether there are segment ordering effects—that is, whether after controlling for all other variables in the SD statistical model shown in Table 5.3, scores recorded in segments occurring earlier in an observation period differ from scores recorded at

[40] Recall that FFT was scored using 30 minute occasions and that too few days had multiple occasions for FFT to be included in the present analysis.

a later point. Note also that in these statistical models, I will report segment timing effects separately for each item on each instrument.

Table 5.4 shows the results for CLASS and Table 5.5 shows the results for PLATO. In both tables, the first column of the table shows the item averages for the reference occasion (i.e. the first occasion; minutes 0-15). The second column shows item-specific deviations from this average due to the second occasion (i.e. minutes 15-30). The third column shows item-specific deviations from this average due to the third occasion (i.e. minutes 30-45). The fourth column shows item-specific deviations from this average due to the fourth and later occasion (i.e. minutes 45 through end of lesson)[41]. All effects are presented in the teacher standard deviation (SD_T) metric.

The results in Tables 5.4 and 5.5 show that observed teaching quality scores varied systematically within the course of a lesson period. In analyses *not* shown here, I found that averaging across all items, scores on CLASS were 0.31 SD_T higher on the second occasion as compared to the first and 0.18 SD_T higher on the third occasion as compared to the first, while the fourth and later occasions were not significantly different than the first occasion. PLATO scores showed a similar effect, but the effect was twice as strong with scores in the second occasion 0.61 SD_T higher and scores on the third occasion 0.40 SD_T higher than scores on the first occasion. This shows that observed teaching quality generally increased through the middle portion of the lesson, remaining lower at the start and end of the lesson.

[41] Only 272 of 901 days (30%) had 4 occasions, 52 (6%) had 5 occasions and 16 (2%) days had six occasions.

Table 5.4: Item-by-Occasion Fixed Effects for the System Design (SD) Model on the CLASS Instrument in the Teacher SD Metric

Item (β_i)	Item Average (Occasion 1)	Occasion 2	Occasion 3	Occasion 4+
Positive Climate	16.87 (0.55)***	0.16 (0.10)	0.03 (0.10)	0.02 (0.15)
Negative Climate	24.62 (0.55)***	0.07 (0.10)	0.16 (0.10)	0.21 (0.15)
Regard for Adolescent Perspectives	11.08 (0.55)***	0.87 (0.10)***	0.85 (0.10)***	1.02 (0.15)***
Teacher Sensitivity	14.81 (0.55)***	0.37 (0.10)***	0.41 (0.10)***	-0.15 (0.15)
Behavior Management	22.09 (0.55)***	-0.14 (0.10)	-0.18 (0.10)	-0.45 (0.15)**
Productivity	21.00 (0.55)***	0.18 (0.10)	0.24 (0.10)*	0.07 (0.15)
Instructional Learning Formats	13.88 (0.55)***	0.23 (0.10)*	-0.28 (0.10)**	-0.97 (0.15)***
Content Understanding	12.25 (0.55)***	0.24 (0.10)*	-0.35 (0.10)***	-1.06 (0.15)***
Analysis and Problem Solving	8.71 (0.55)***	0.64 (0.10)***	0.63 (0.10)***	0.49 (0.15)**
Quality of Feedback	12.16 (0.55)***	0.65 (0.10)***	0.58 (0.10)***	0.37 (0.15)*
Student Engagement	18.49 (0.55)***	0.13 (0.10)	0.07 (0.10)	0.09 (0.15)

Note. Column 'Main' shows the Item mean on occasion 1; Column '2' shows the deviation of the item on occasion 2; Column '3' shows the deviation of the item on occasion 3; Column '4+' shows the deviation of the item on occasion 4 or higher. * p<0.05; ** p<0.01; *** p<0.001.

Table 5.5: Item-by-Occasion Fixed Effects for the System Design (SD) Model on the PLATO Instrument in the Teacher SD Metric

Item (β_i)	Item Average (Occasion 1)	Occasion 2	Occasion 3	Occasion 4+
Purpose	27.25 (0.62)***	0.08 (0.19)	-0.04 (0.20)	-0.29 (0.30)
Intellectual Challenge	19.42 (0.62)***	1.13 (0.19)***	1.33 (0.20)***	1.07 (0.30)***
Representation of Content	22.77 (0.62)***	0.75 (0.19)***	0.18 (0.20)	-0.93 (0.30)**
Connections to Prior Knowledge	16.06 (0.62)***	-1.48 (0.19)***	-2.93 (0.20)***	-4.33 (0.30)***
Connections to Personal Experience	12.60 (0.62)***	0.40 (0.19)*	-0.08 (0.20)	-0.47 (0.30)
Explicit Strategy Instruction	11.48 (0.62)***	0.14 (0.19)	-0.29 (0.20)	-0.51 (0.30)
Modeling	11.62 (0.62)***	0.91 (0.19)***	0.51 (0.20)*	0.10 (0.30)
Guided Practice	21.84 (0.62)***	1.24 (0.19)***	2.29 (0.20)***	2.24 (0.30)***
Classroom Discourse	19.02 (0.62)***	1.32 (0.19)***	1.01 (0.20)***	0.71 (0.30)*
Text Based Instruction	16.53 (0.62)***	2.71 (0.19)***	3.08 (0.20)***	3.03 (0.30)***
Acc. for Language Learning	13.34 (0.62)***	0.15 (0.19)	-0.43 (0.20)*	-1.06 (0.30)***
Behavior Management	37.00 (0.62)***	-0.30 (0.19)	-0.27 (0.20)	-0.20 (0.30)
Time Management	34.86 (0.62)***	0.86 (0.19)***	1.08 (0.20)***	1.28 (0.30)***

Note. Column 'Main' shows the Item mean on occasion 1; Column '2' shows the deviation of the item on occasion 2; Column '3' shows the deviation of the item on occasion 3; Column '4+' shows the deviation of the item on occasion 4 or higher; Acc. for Language Learn is Accommodations for Language Learning. * p<0.05; ** p<0.01; *** p<0.001.

This average effect, however, hides the large heterogeneity of occasion effects across items, as shown in Tables 5.4 and 5.5. The item-specific effects, when statistically significant, were often much larger than the item average effects just reported, but these effects also varied widely across items. The patterns of occasion effects on specific items defy easy description. The results shown for CLASS in Table 5.4, for example, show that scores on the Regard for Adolescent Perspectives, Teacher Sensitivity, Analysis and Problem Solving, and Quality of Feedback items generally increased in later lesson occasions compared to the first occasion; that Positive Climate, Negative Climate, and Student Engagement scores were steady across lesson occasions; that scores on Instructional Learning Formats and Content understanding peaked in the second lesson occasion and decline thereafter; and that scores on Behavior Management declined at the end of lessons.

Table 5.5 shows item-specific effects of occasion on PLATO item scores. Here, the data showed that scores on Connections to Prior Knowledge was highest at the very beginning of lessons; that scores on Intellectual Challenge, Modeling, Guided Practice, Classroom Discourse, Text-based Instruction, and Time Management generally increased after the first lesson occasion; that Purpose, Explicit Strategy Instruction, and Behavior Management remained constant across the lesson; and that Representations of Content, Accommodations for Language Learning, and Connections to Personal Experience were higher in the second occasion before falling off towards the end of the lesson. Taken as a whole, then, the results in Tables 5.4 and 5.5 suggest that occasion order is indeed a hidden facet that has effects on observed teaching quality. Thus, if observations were only conducted on some occasions within a lesson, as they often are during informal observations (Steinberg & Donaldson, 2015), estimates of teacher quality and the item-specific feedback received by teachers will vary depending on the occasions sampled. Ratings will be higher when the middle of a day of instruction is sampled rather than the start or end of the lesson.

In summary, then, the SD model just discussed shows that a variety of SD facets affect observed teaching quality scores. Observed teaching quality is dependent on when teachers are observed, whether scoring was done live or from video, and when raters did the scoring. The analyses presented here further suggest that these facet effects can be large enough, especially if examined in combination, to have important effects on where teachers fall in the distribution of teacher quality. Importantly, while the analyses here show that SD facets affected observed teaching quality ($X_{ir(o:d:s:t)}$), it is not clear how much they affect estimates of a teacher's teacher quality ($\hat{\upsilon}_t$). This was because the design of UTQ controlled for the impact of these facets by sampling days across the full school year, by randomly assigning raters to observation days, and by randomly ordering dates of scoring. Each of these steps balances the impact of these facets across the four days each teacher was observed, helping to minimize their effect on estimates of teacher quality ($\hat{\upsilon}_t$). Below, I will show that the estimated variance of teacher quality ($var(\hat{v}_t^{Base})$) in the Base model was inflated compared to the estimated variance of teacher quality ($var(\hat{v}_t^{SD})$) in the SD model, but the estimated teacher quality scores from the Base and SD model were almost identical ($cor(v_t^{Base}, v_t^{SD}) \approx 1$). This demonstrates that not controlling for the SD facets inflated the estimate of the variance of teacher quality ($var(\hat{v}_t^{Base})$), but had little effect on the estimates themselves ($\hat{\upsilon}_t^{Base}$), at least in UTQ data where sampling was well controlled.

V.2.2. Curriculum and Instruction Model (CI) I turn now to reporting the results of the Curriculum and Instruction (CI) model, which adds effects for the CI facets to the SD model just discussed[42]. Like the last section, the goal of this section is to show that the CI facets have effects on observed teaching quality (i.e. $X_{ir(o:d:s:t)}$). The CI facets are analyzed here to investigate whether observed teaching quality is affected by various characteristics of the content being taught in a lesson and the structure framing the interactions between students and teachers. These facets have the potential to cause bias across instruments (RQ2a) because instruments may favor types of instructional approaches

[42] The effects of curriculum and instructional facets are somewhat affected by how the facets are created. Appendix E shows the results for the CI and SO model when hidden facets are created through averaging PLATO log items across segments and raters. All hidden facets are positively related to observed teaching quality on PLATO under this construction, the effect of literature and grammar are positive on CLASS, and the effect of literature is non-significant on FFT. These changes are more the result of shifts across the p=0.05 threshold rather than large changes in the parameters themselves.

that are not *always* ideal. Further, when these facets are hidden (i.e. excluded from the statistical model), they are likely to inflate estimates of the variance of teacher quality ($\widehat{var}(v_t)$), as I argued earlier and demonstrate statistically in future sections. Here, I focus on whether the hidden facets affect observed teaching quality ($X_{ir(o:d:s:t)}$), which is a necessary pre-requisite for them to affect estimates of teacher quality ($\hat{v}_t$). I also focus here on whether hidden facet effects are consistent across instruments, as inconsistency in effects necessarily implies instrument bias.

Table 5.6 shows the results of the CI model where effects are reported in terms of the teacher standard deviation metric with standard errors in parentheses. Each column in the table presents the fixed effect estimates from the instrument-specific GTheory regression model. I focus first on the results for content domain facets. Note that the results for the SD facet effects already discussed do not change much (except for the effect of double scoring on PLATO as discussed in footnote 39).

Table 5.6: Fixed Effects for the Curriculum and Instruction (CI) Model across the three Instruments in the Teacher SD Metric

Names	CLASS	FFT	PLATO
Scored Live (β_{Live})	0.37 (0.20)	0.54 (0.21)**	-0.40 (0.24)
Double Scored (β_{Dbl})	-0.11 (0.13)	0.06 (0.13)	-0.43 (0.14)**
Date Scored (m) (β_{DtSc})	-0.06 (0.01)***	-0.04 (0.01)**	-0.06 (0.02)***
Day of the Week (β_{DayWk})			
Tuesday	-0.03 (0.15)	0.06 (0.14)	-0.03 (0.17)
Wednesday	0.32 (0.16)*	0.32 (0.15)*	0.11 (0.18)
Thursday	-0.01 (0.15)	0.19 (0.15)	-0.17 (0.17)
Friday	-0.27 (0.18)	-0.16 (0.18)	-0.07 (0.21)
Observation Month (β_{Month})	-0.11 (0.02)***	-0.09 (0.02)***	-0.07 (0.03)*
Content Domain			
Reading (β_{Read})	0.09 (0.19)	-0.26 (0.18)	0.48 (0.22)*
Literature (β_{Lit})	0.36 (0.14)**	0.44 (0.13)***	1.08 (0.16)***
Writing (β_{Write})	0.43 (0.13)***	0.21 (0.12)	0.99 (0.15)***
Grammar ($\beta_{Grammar}$)	0.17 (0.13)	-0.25 (0.13)*	0.02 (0.15)
Interaction Structure			
Discussion (β_{Disc})	0.28 (0.10)**	0.03 (0.10)	0.67 (0.12)***
Independent (β_{Ind})	0.05 (0.17)	0.21 (0.16)	0.44 (0.20)*
Recitation (β_{Rec})	-0.17 (0.11)	-0.03 (0.11)	0.21 (0.15)

Note. Each column shows the results of a separate model for the indicated instrument. Date Scored is scaled so a 1 point difference is one month; Monday is the reference group for the Days of the Week. * p<0.05; ** p<0.01; *** p<0.001.

As Table 5.6 shows, observed teaching quality (i.e. $X_{ir(o:d:s:t)}$) did indeed vary as a result of whether or not there was a sustained focus on a content domain of interest[43,44]. The most consistent finding in the table was that lessons with a sustained focus on literature generally received higher scores on all instruments compared to lessons with no sustained content focus (ES=0.36 SD_T for CLASS, 0.44 SD_T for FFT, and 1.08 SD_T for PLATO). For two instruments (CLASS and PLATO) lessons with a sustained focus on writing also received higher scores than lessons having no sustained content focus (ES=0.43 SD_T for CLASS and 0.99 SD_T for PLATO). These effects are capable of moving a teacher from the 50th percentile of teacher quality to the 67th percentile on CLASS scores and the 84th

[43] Note that there is technically no reference group because these facets are not mutually exclusive. The correct interpretation is the effect of the facet compared to lessons not using that facet. However, there is relatively little overlap in content domains so the "reference group", in effect, is a set of lessons that have no sustained focus on any of the four focal content domains (roughly 1/4 of lessons). This includes both lessons that change domain across occasions and those that never focus on a domain. This makes the "reference group" difficult to clearly conceptualize.

[44] The reader should notice that effects are generally slightly larger and present more often for PLATO. This is likely due to the same rater providing PLATO scores and the PLATO log scores, but may also reflect the PLATO instrument's special sensitivity to specific types of instruction. The data does not allow me to separate these possibilities.

percentile on PLATO scores. Further, PLATO scores were 0.48 SD_T higher on lessons that focus on reading and FFT scores were 0.25 SD_T lower for lessons that focus on grammar[45]. Past research by (Grossman et al., 2014) showed a similar negative effect of grammar lessons on PLATO relative to lessons focused on reading and writing lessons. This result is confirmed in the UTQ data and was extended to the FFT instrument (Note that PLATO scores on grammar lessons were significantly lower than PLATO scores on literature and writing lessons though not lower than scores on lessons with no sustained focus on a content domain; see footnote 45).

Table 5.6 also shows the effects on observed teaching quality (i.e. $X_{ir(o:d:s:t)}$) from hidden facets involving classroom interaction structures (i.e. whether lessons included discussion, recitation, and independent work). The interaction structure of a lesson had some effects on observed teaching quality, though the effects differed across instruments. PLATO scores increased when lessons included discussion and independent work (ES=0.67 SD_T and 0.44 SD_T respectively); CLASS scores only increased when lessons included discussion (ES=0.28 SD_T); and there were no classroom interaction facet effects on FFT scores. Overall, then, both content domain and interaction structure can be considered hidden facets of measurement. Since a teacher is likely to be observed across multiple levels of these facets, they should be considered within-teacher facets and should act to reduce the precision of measurement and inflate the estimate of the variance of teacher quality ($\hat{var}(v_t)$). This does not, however, mean there are no differences between teachers in the frequency of instruction across levels of the content domain or interaction structure facets, which could lead to

[45] In analyses not shown here, I ran contrast tests to explore whether the curriculum effects on observed teaching quality just discussed differed across the four content domains. On the CLASS instrument, the effects of the four content domains were not distinguishable from each other. Thus, while observed CLASS scores for some content domains differed from lessons with no sustained focus on a content domain, lessons with a sustained focus on a content domain did not differ from each other. On FFT and PLATO, literature and writing lessons received higher scores than grammar lessons. Further, on FFT only, literature and writing lessons also received higher scores than reading lessons.

between-teacher effects of this facet (and potentially bias), a question I return to in a later section.

Table 5.6 additionally shows that the effects of specific content domains and interaction structures varied across the three instruments, a sign of instrument bias. Recall that this is a sign of instrument bias because the true teaching quality (i.e. X_{true}) cannot be both higher and lower *nor* higher and not higher at the same time (see Appendix F for a more detailed discussion of the determination of instrument bias). While instrument bias can be see by examining the different point estimates of the CI facets in Table 5.6, Figure 5.1 presents a more clear view of the problem of instrument bias. Figure 5.1 shows the range of point estimates obtained for the effect of the hidden facet on observed teaching quality across bootstrap replicates, along with means and 95% confidence intervals for those estimates. Under the parametric bootstrap assumptions and the null-hypothesis of no instrument bias, the distributions of these point estimates should overlap. To the extent that they do not, there is evidence of instrument bias (consult Appendix F for a more detailed discussion of this point). From Figure 5.1, we can see that the effect of reading on PLATO scores is much larger than the effect on FFT scores (p=0.008); the effect of literature on PLATO scores is much larger than the effect on FFT scores (p<0.001) and CLASS scores (p<0.001); the effect of writing on PLATO scores is much larger than the effect on FFT scores (p<0.001) and CLASS scores (p=0.002); and the effect of grammar on FFT scores is less than the effect on CLASS scores (p=0.010). Further, there are also differences in the size of the effect of interaction structure facets across the instruments. The effect of discussion lessons on PLATO scores is significantly greater than the effect on CLASS scores (p=0.008) and FFT scores (p<0.001). Thus, there is some evidence of instrument bias for all the content domain facets and for discussion lessons with the main effect being the effect on PLATO scores was much larger than the effect on the other instruments.

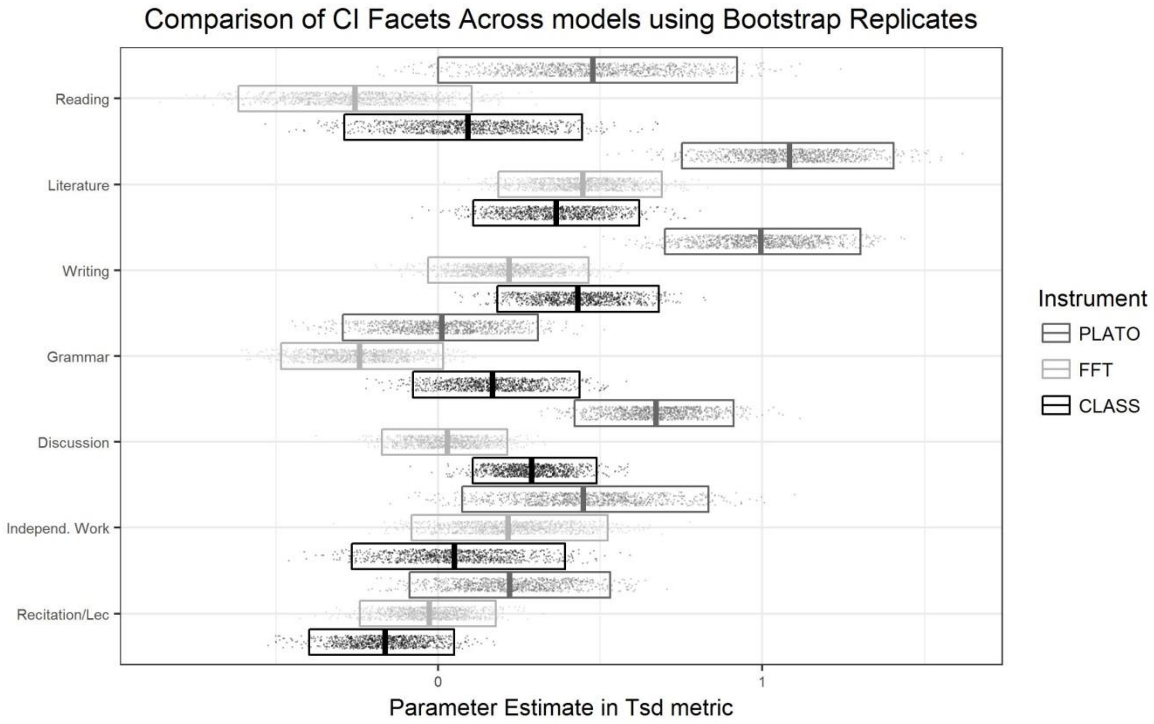

Figure 5.1: Comparison of CI Facet Effects across Bootstrap Replicates

As I have argued, the most likely sources of this bias are either construct under-representation or construct-irrelevant variance. Given the prominent role that PLATO plays in the findings, the instrument bias appears to be driven by differences between PLATO and the other instruments. Being the only subject-specific instrument, PLATO scores may be more sensitive to the aspects of instruction that vary across the content domain being taught and interaction structure being used. That is, PLATO scores measure some important aspect of teaching quality or measure some irrelevant feature of instruction unrelated to teaching quality that varies across the facets examined. I explored these possibilities using the item-specific models presented in Appendix D. The positive effect of reading lessons on PLATO scores arose from positive effects on items related to text-based instruction, explicit strategy instruction, and accommodations for language learners; dimensions of instruction not captured by FFT. The negative effect for FFT was due to questioning and discussion techniques, engaging students in learning, and using assessment in instruction. Of these FFT dimensions, only questioning and discussion techniques was captured directly by PLATO (and the effect of classroom discussion on PLATO was slightly negative). This suggests that reading lessons were more centered on texts and explicitly introducing reading strategies with some recognition of language learners, while not including discussions or assessment and were less engaging.

Common patterns of item effects were also visible for the effect of literature. Across the three instruments, no effect was significant for items related to behavior management and time management while the effect of being a literature lesson was positive on items related to instructional quality. The larger effect on PLATO scores, then, seems the result of the greater focus on instructional items and lesser focus on classroom culture, with the PLATO focus on text-based instruction playing a prominent role in explaining why PLATO scores were much more strongly related to literature lessons than the other two instruments. The negative effect of grammar lessons on FFT scores stemmed from lower scores on FFT's culture of learning and questioning and discussion items, though most items had a non-significant negative coefficient. This suggests grammar lessons lacked academic press (Shouse, 1996) and discussions, which are not directly scored by CLASS.

When lessons involved discussion, PLATO scores were higher on all items, except modeling, accommodations for language learners, behavior management, and time management while, for FFT, only organizing physical space and questioning and discussion techniques received higher scores on discussion lessons. This suggests discussion lessons included discussions and included a number of English specific beneficial strategies (i.e. PLATO only items), but had weak time management, behavior management, and modeling of strategies. Thus, for the cases of instrument bias identified in the CI facets , there was evidence that the bias between instruments was a function of the specific dimensions of instruction being measured, as effects were isolated to dimensions only measured well by one instrument. However, an alternative explanation cannot be ruled out. The same rater provided PLATO scores and PLATO log information that was used to create the CI facets. This could lead to correlated error in the PLATO scores and PLATO log due to the common rater, which may also explained the observed larger effects on PLATO scores than on CLASS and FFT scores. There is no way to rule out this alternative explanation. Only the instrument bias in grammar scores, which did not involve the PLATO instrument, is free of the contamination of rater error.

Overall, I have provided evidence that demonstrates that, at least in UTQ data, the content domain being taught and the interaction structure of the lesson both impacted observed teaching quality (i.e. $X_{ir(o:d:s:t)}$). As I have argued, this should lead models that do not account for these effects to be less precise, with inflated estimates of the variance of teacher quality ($\widehat{var}(v_t)$). I will explore the exact extent of this effect below, as well as test for between-teacher effects of the CI facets that may introduce bias into estimates of teacher quality ($\widehat{v_t}$). I also showed in this section that there is some indication of instrument bias for reading lessons, literature lessons, writing lessons, grammar lessons, and discussion lessons. This bias seems to stem from the specific aspects of teacher quality that each instrument measures (i.e. producing either construct under-representation or construct-irrelevant variance), though it may be due to rater error. When instruments contain different items and when only some aspects of instruction change across levels the CI facets can adopt (e.g. reading or discussions), then only those instruments with items capturing aspects of teaching that differ across levels of the CI facets will show differences in observed scores across these facets. Adopting a statistical model (such as the CI model) that controls for the CI facets will ensure that the instrument biases identified here do not contaminate estimates of teacher quality ($\widehat{v_t}$), at least when the bias takes the form of an average difference in observed scores. Only scores for the instrument showing bias would have to be adjusted, but, as I have discussed, it is not possible to identify which specific instruments showed bias.

V.2.3. School Organization Model (SO) In this section, I describe the results for the School Organization (SO) model. As in the previous two prior hidden facet models (above), the goal of this section is to examine the effects that SO facets have on observed teaching quality (e.g. $X_{ir(o:d:s:t)}$). I also will look for evidence of instrument bias to address RQ 1a.

Table 5.7 shows the results of this analysis. The model estimated here simply added the SO variables to the CI statistical model just discussed. The variables added to the model are: grade level of the class section being taught, the average prior year's achievement level of students in a class section, and the average score of students in a class section on the demographic composite discussed earlier. Once again, the effects of these variables on observed scores are presented separately for each observation instrument, all effects are reported in the SD_T metric, and standard errors are in parentheses. In the discussion, I will focus only on the effects of the School Organization facets. However, it is worth noting that in contrast to results for the CI model discussed above, adding School Organization variables to the model changes some estimates for other variables. The most notable change is that the estimated effect of literature lessons on observed scores decreases in this new model, suggesting that (as I found in analyses not presented here) at least part of the literature effect on teaching quality comes from differences in student background across classes, where classrooms having higher achieving students also are more likely to teach literature.

Table 5.7: Fixed Effects for School Organization (SO) Model in Teacher SD Metric

Names	CLASS	FFT	PLATO
Scored Live (β_{Live})	0.48 (0.20)*	0.65 (0.20)**	-0.30 (0.24)
Double Scored (β_{Dbl})	-0.16 (0.13)	0.03 (0.13)	-0.46 (0.14)**
Date Scored (m) (β_{DtSc})	-0.04 (0.01)***	-0.03 (0.01)*	-0.05 (0.02)***
Day of the Week (β_{DayWk})			
Tuesday	-0.06 (0.14)	-0.02 (0.14)	-0.10 (0.17)
Wednesday	0.24 (0.15)	0.25 (0.15)	0.03 (0.18)
Thursday	-0.04 (0.14)	0.11 (0.14)	-0.25 (0.17)
Friday	-0.26 (0.18)	-0.21 (0.17)	-0.10 (0.21)
Observation Month (β_{Month})	-0.11 (0.02)***	-0.09 (0.02)***	-0.07 (0.03)**
Content Domain			
Reading (β_{Read})	0.16 (0.18)	-0.16 (0.18)	0.55 (0.21)*
Literature (β_{Lit})	0.25 (0.13)	0.30 (0.13)*	0.96 (0.16)***
Writing (β_{Write})	0.42 (0.12)***	0.18 (0.12)	0.98 (0.15)***
Grammar ($\beta_{Grammar}$)	0.17 (0.13)	-0.24 (0.12)*	0.03 (0.15)
Interaction Structure			
Discussion (β_{Disc})	0.23 (0.10)*	-0.04 (0.10)	0.64 (0.12)***
Independent (β_{Ind})	0.05 (0.16)	0.20 (0.15)	0.44 (0.19)*
Recitation (β_{Rec})	-0.16 (0.11)	-0.00 (0.11)	0.22 (0.15)
Grade			
7th Grade (β_{7th})	-0.51 (0.16)**	-0.37 (0.15)*	-0.42 (0.18)*
8th Grade (β_{8th})	0.07 (0.15)	0.07 (0.15)	0.13 (0.17)
Prior Achievement (β_{PrAch})	0.29 (0.09)**	0.44 (0.08)***	0.17 (0.10)
St. Info Missing (β_{Imp})	-0.48 (0.25)	-0.39 (0.24)	-0.32 (0.28)
Demographic Composite (β_{Demo})	-0.33 (0.09)***	-0.20 (0.09)*	-0.27 (0.10)**

Note. Each column shows the results of a separate model for the indicated instrument. Date Scored is scaled so a 1 point difference is one month. Monday is the reference group for the Days of the Week. Prior Achievement is captured at the section level and is the average achievement level on last year's standardized test for students in a particular section. The Demographic Composite is a section-level variable and captures classrooms that have higher percentages of students who are black, Hispanic, ELL, and FRL. St. Info Missing is a dummy variable indicating if Prior Achievement and Demographic Composite are missing. * p<0.05; ** p<0.01; *** p<0.001.

The first variable to be discussed is grade level. As Table 5.7 shows, seventh grade classrooms receive lower scores than 6th and 8th grade classrooms. On CLASS, seventh grade scores are 0.51 SD$_T$ lower than 6th grade classrooms; on FFT, scores are 0.37 SD$_T$ lower than 6th grade classrooms; and on PLATO, seventh grade scores are 0.42 SD$_T$ lower than 6th grade classrooms. These findings do *not* match past work of Grossman and colleagues (2014), who did not find grade level effects within middle schools (but did find middle schools received lower scores than elementary schools).

Table 5.7 also shows large effects on observed teaching quality of my two measures of student characteristics in the class sections taught by UTQ teachers. Across all instruments, class sections with more disadvantaged students received lower observation scores than those with fewer disadvantaged students. In fact, for every standard deviation increase in the demographic composite (which represents classrooms becoming more black, Hispanic, ELL, and FRL), there was a 0.33 SD_T decrease in CLASS scores, a 0.20 SD_T decrease in FFT scores, and a 0.27 SD_T decrease in PLATO scores. Additionally, on CLASS and FFT, class sections with lower average prior achievement scores received lower observation scores. Here, for every standard deviation increase in the section average student prior achievement, there was an increase of 0.29 SD_T in CLASS, an increase of 0.44 SD_T in FFT scores, and a statistically insignificant increase of 0.17 SD_T in PLATO scores. These results are reasonably consistent across instruments, but smaller for PLATO scores than for CLASS and FFT scores for students' prior achievement (the smaller effect on PLATO scores is not significant as can be seen in Appendix F). Past work has suggested that instruction becomes more controlling and directed when there are more disadvantaged students in a classroom (Carlisle et al., 2011), but the item specific models in Appendix D provide neither clear evidence to confirm this possibility nor suggest another explanation of these results.

The results from the SO model are similar to results from the models presented in previous sections, providing strong evidence the SO facets are hidden facets. This is important because it suggests that these SO effects can affect generalizations and/or extrapolations researchers might want to make to a pre-defined universe. In earlier discussions of these SO variables, I have hypothesized that SO effects can be driven by either "co-construction" effects (where teachers and students jointly produce quality of teaching), teacher sorting effects (where better teachers tend to work in more advantaged classrooms and schools), or other unknown effects. The adjustments presented in the SO model are only appropriate if co-construction (or a similar effect) drive differences in observed teaching quality. If teacher sorting explains the facet effects, then estimates of teacher quality from the SO model are likely to be biased. Interestingly, the data presented in Table 5.7 provide no evidence of instrument. In the next section, I explore whether these SO facet effects were associated with within-teacher, between-teacher, or between-school differences in observed teaching quality, which has important implications for how the SO facets might affect estimates of teacher quality.

V.2.4. Within-Teacher and Within-School Effects In the previous sections, I demonstrated that some facets in each of the three categories of hidden facets were systematically associated with variations in observed teaching quality ($X_{ir(o:d:s:t)}$). By definition, this makes them "hidden facets" of measurement. As I have argued, one of the main determinants for how a hidden facet affects estimates of teacher quality ($\widehat{\upsilon_t}$) is whether the effect of the hidden facet on observed teaching quality ($X_{ir(o:d:s:t)}$) acts within-teachers or between-teachers. Further, identifying between-school differences is important because isolating facet effects from broader school context effects is difficult. Hidden facets that act within-teachers are likely to inflate estimates of the variance of teacher quality ($\widehat{var(\upsilon_t})$), but not cause bias to estimates of teacher quality ($\widehat{\upsilon_t}$). Hidden facets that act between-teachers, including those acting between schools, may lead to bias in estimates of teacher quality ($\widehat{\upsilon_t}$) because co-construction-like effects imply differences in observed teaching quality ($X_{ir(o:d:s:t)}$) that are not solely the result of differences in teacher knowledge or ability. In this section, I test to see the level of nesting (i.e. within-teacher, between-teacher, or between-schools) at which hidden facets affected observed teaching quality.

Tables 5.8, 5.9, and 5.10 show, for CLASS, FFT, and PLATO, respectively, the results from statistical models based on the SO model. Note these tables separate the CI and SO facets into three components: a within-teacher component, a between-teacher component, and a between-schools component. I only show those CI and SO facets that could act either within-teachers, between-teachers, or between-schools (i.e. I do not include grade). Tables 5.8, 5.9, and 5.10 present the original SO model estimates in the left column. The next three columns come from a single model, breaking the facets down into a within-teacher component (second column), a between-teacher, within-school component (third column), and a between-school component (last column).

Before discussing the results of Tables 5.8, 5.9, and 5.10, I first want to bring attention to the standard errors of estimates across the three components. The content domain and interaction structure standard errors were very large between-teachers and between-schools, which reflects the fact that most of the variation in these facets was within-teachers (i.e. there was little variation of the CI facet variables between-teachers and between-schools). The prior achievement standard errors were about equal across components, though up to twice as large as when including only the single prior achievement effect. The standard errors on the demographic composite parameters were smallest for between-school variation, reflecting the large between-school variation of this facet (over 80% of the variance in demographic composite was between schools). Thus, I have the most power to detect within-teacher effects from the CI facets and between-school effects for the demographic composite facet while I have about equal power for each component of the prior achievement facet.

Table 5.8: Within-Teacher, Between-Teacher, and Between-School Effects on Observed Teaching Quality of the CI and SO facets for CLASS

Facet	SO Model	Within-Teacher	Between-Teacher	Between-School
Content Domain				
Reading (β_{Read})	0.16 (0.18)	0.30 (0.20)	-0.58 (0.51)	-1.02 (1.06)
Literature (β_{Lit})	0.25 (0.13)	0.14 (0.15)	0.78 (0.36)*	0.97 (0.68)
Writing (β_{Write})	0.42 (0.12)***	0.52 (0.14)***	0.17 (0.36)	-0.67 (0.76)
Grammar ($\beta_{Grammar}$)	0.17 (0.13)	0.29 (0.14)*	-0.20 (0.33)	-0.63 (0.73)
Interaction Structure				
Discussion (β_{Disc})	0.23 (0.10)*	0.13 (0.11)	0.60 (0.32)	1.37 (0.68)*
Independent (β_{Ind})	0.05 (0.16)	0.03 (0.17)	0.27 (0.49)	-0.05 (0.93)
Recitation (β_{Rec})	-0.16 (0.11)	-0.16 (0.12)	-0.48 (0.38)	-0.39 (0.76)
Prior Achievement (β_{PrAch})	0.29 (0.09)**	0.13 (0.15)	0.16 (0.14)	0.39 (0.19)*
Demographic Composite (β_{Demo})	-0.33 (0.09)***	-0.26 (0.26)	-0.64 (0.28)*	-0.07 (0.14)

Note. The model included all of the parameters from the SO model, but only the CI and SO facets are displayed. The left column presents the parameter estimates from the original SO model. The other three columns show results from a single model. The next column presents the within-teacher component, estimated as the original value minus the teacher mean score. The third column presents the between-teacher component, estimated as the teacher mean score minus the school mean score. The last column presents the between-school component, estimated as the mean aggregated to the school level. The effects are in the teacher quality standard deviation metric (SD_T). The Demographic Composite represents classrooms that have higher percentages of students who are black, Hispanic, ELL, and FRL. * p<0.05; ** p<0.01; *** p<0.001.

Table 5.9: Within-Teacher, Between-Teacher, and Between-School Effects on Observed Teaching of the CI and SO facets for FFT

Facet	SO Model	Within-Teacher	Between-Teacher	Between-School
Content Domain				
Reading (β_{Read})	-0.16 (0.18)	-0.17 (0.19)	-0.32 (0.49)	-0.05 (1.04)
Literature (β_{Lit})	0.30 (0.13)*	0.29 (0.14)*	0.38 (0.36)	-0.02 (0.67)
Writing (β_{Write})	0.18 (0.12)	0.25 (0.13)*	-0.01 (0.35)	-0.76 (0.75)
Grammar ($\beta_{Grammar}$)	-0.24 (0.12)*	-0.18 (0.13)	-0.33 (0.33)	-0.98 (0.72)
Interaction Structure				
Discussion (β_{Disc})	-0.04 (0.10)	-0.11 (0.10)	0.42 (0.31)	0.89 (0.67)
Independent (β_{Ind})	0.20 (0.15)	0.18 (0.16)	0.25 (0.48)	0.62 (0.91)
Recitation (β_{Rec})	-0.00 (0.11)	-0.07 (0.11)	0.23 (0.37)	0.05 (0.74)
Prior Achievement (β_{PrAch})	0.44 (0.08)***	0.18 (0.15)	0.50 (0.14)***	0.55 (0.18)**
Demographic Composite (β_{Demo})	-0.20 (0.09)*	-0.24 (0.25)	-0.36 (0.28)	-0.05 (0.14)

Note. The model included all of the parameters from the SO model, but only the CI and SO facets are displayed. The left column presents the parameter estimates from the original SO model. The other three columns show results from a single model. The next column presents the within-teacher component, estimated as the original value minus the teacher mean score. The third column presents the between-teacher component, estimated as the teacher mean score minus the school mean score. The last column presents the between-school component, estimated as the mean aggregated to the school level. The effects are in the teacher quality standard deviation metric (SD_T). The Demographic Composite represents classrooms that have higher percentages of students who are black, Hispanic, ELL, and FRL. * p<0.05; ** p<0.01; *** p<0.001.

Table 5.10: Within-Teacher, Between-Teacher, and Between-School Effects on Observed Teaching of the CI and SO facets for PLATO

Facet	SO Model	Within-Teacher	Between-Teacher	Between-School
Content Domain				
Reading (β_{Read})	0.55 (0.21)*	0.64 (0.24)**	0.09 (0.56)	0.54 (1.17)
Literature (β_{Lit})	0.96 (0.16)***	0.91 (0.18)***	1.37 (0.41)***	1.09 (0.75)
Writing (β_{Write})	0.98 (0.15)***	1.01 (0.16)***	1.26 (0.40)**	0.07 (0.85)
Grammar ($\beta_{Grammar}$)	0.03 (0.15)	0.12 (0.17)	-0.19 (0.37)	-0.16 (0.81)
Interaction Structure				
Discussion (β_{Disc})	0.64 (0.12)***	0.47 (0.13)***	1.37 (0.36)***	2.60 (0.76)***
Independent (β_{Ind})	0.44 (0.19)*	0.45 (0.21)*	0.21 (0.54)	0.23 (1.03)
Recitation (β_{Rec})	0.22 (0.15)	0.24 (0.16)	0.06 (0.42)	-0.63 (0.84)
Prior Achievement (β_{PrAch})	0.17 (0.10)	0.12 (0.18)	-0.10 (0.16)	0.27 (0.21)
Demographic Composite (β_{Demo})	-0.27 (0.10)**	-0.21 (0.30)	-0.89 (0.32)**	-0.01 (0.16)

Note. The model included all of the parameters from the SO model, but only the CI and SO facets are displayed. The left column presents the parameter estimates from the original SO model. The other three columns show results from a single model. The next column presents the within-teacher component, estimated as the original value minus the teacher mean score. The third column presents the between-teacher component, estimated as the teacher mean score minus the school mean score. The last column presents the between-school component, estimated as the mean aggregated to the school level. The effects are in the teacher quality standard deviation metric (SD_T). The Demographic Composite represents classrooms that have higher percentages of students who are black, Hispanic, ELL, and FRL. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

Table 5.8 shows that, for the CLASS instrument, literature lessons (ES=0.78 SD_T) affected between-teacher within-school differences in observed teaching quality while writing (ES=0.52 SD_T) and grammar (ES =0.29 SD_T) lessons affected within-teacher between-day differences in teaching quality. Additionally, discussion lessons affected between-school differences in observed teaching quality (ES=1.37 SD_T). This means that the average teacher, when observed on CLASS during a writing or grammar lesson, scored higher (on average) than when observed during a non-writing or non-grammar lessons. Additionally, teachers who were observed teaching literature more often than other teachers in their school had higher average observed teaching quality than those teachers who were observed teaching literature less often than other teachers in their school. Last, schools where discussion lessons were more commonly observed had higher average observed teaching quality than schools where discussion lessons were less common. This is surprising because, as I have argued, the CI facets should mostly affect observed teaching quality within-teachers, between-days since all teachers will engage in instruction across the range of content domains and interaction structures during the observation period. In fact, over 85% of the variance in the frequency of CI facets occurred within-teachers (which is why the standard errors for the within-teacher component are much lower).

As Table 5.9 shows, on FFT, no CI facets affected between-teacher or between-school differences in observed teaching quality. The literature (ES=0.29 SD_T) and writing (ES=0.25 SD_T) lesson effects, as predicted, affected within-teacher, between-day differences in observed teaching quality. Table 5.10 shows that PLATO was more similar to CLASS than FFT, though far more effects on PLATO were significant as compared to the other two instruments. Scores on reading (ES=0.64 SD_T) and independent work (ES=0.45 SD_T) lessons were higher within-teachers between-days than scores on lessons with no sustained focus on those facets. Further, Table 5.10 shows that literature and writing lessons were associated with higher observed teaching quality within-teachers between-days (ES=0.91 SD_T and ES=1.01 SD_T respectively) *and* between-teachers within-schools (ES=1.37 SD_T and ES=1.26 SD_T respectively) while discussion lessons had higher observed teaching quality across all three components (ES=0.47 SD_T within-teachers; ES=1.37 SD_T between-teachers; and ES=2.60 SD_T between-schools). That is, the average teacher had higher observed teaching quality on PLATO during discussion lessons than that same teacher received on non-discussion lessons *and* teachers who were observed teaching more discussion lessons had higher average observed teaching quality on PLATO than those observed teaching fewer discussion lesson *and* schools where teachers were observed teaching more discussion lessons received higher average observed teaching quality on PLATO than schools where teachers were observed teaching fewer discussion lessons. One caveat to these findings is that the discrepancies across instruments (e.g. only within-teacher writing effects are significant across all three instruments) call into question the stability of these effects. Note that the same instrument biases on CI facets identified earlier occurred here, but only within-teachers. There was not enough precision in the between-teacher and between school effects to identify differences across instruments.

Contrary to expectations, Tables 5.8, 5.9, and 5.10 show that, while most effects of CI facets are within-teachers, there are some between teacher effects of CI facets, at least for CLASS and PLATO. I turn here to break down the implications of the different components shown in Tables 5.8, 5.9, and 5.10. The within-teacher component captures differences in observed teaching quality for a teacher when she/he, for example, teaches writing lessons compared to when she/he teaches non-writing lessons. When the within-teacher component of a CI facet was significant (e.g. writing across all three instruments), estimates of models that do not control for the CI facets, as I have argued before, will have inflated estimates of the variance of teacher quality ($\widehat{var}(\upsilon_t)$) and less precise estimates of teacher quality ($\widehat{\upsilon_t}$) because average teacher scores will vary randomly due to the sampling of days within-teachers. Then, to increase the precision of measuring teacher quality, we should prefer the CI model that adjusts for these facets and removes this source of imprecision. When the between-teacher component of a CI facet is significant (e.g. literature on CLASS), it implies teacher sorting is occurring (e.g. teachers with higher average scores on CLASS were observed teaching more literature lessons) *because* the within-teacher component controls for any possible co-construction effects (e.g. if it were easier to enact high quality teaching on literature lessons [i.e. co-construction], the within-teacher component would adjust for this, assuming there is sufficient within-teacher variation in these facets, as is the case in the UTQ data) [46]. It is this between-teacher effect that captures differences in teacher quality due to

[46] Note that this is a variation on what I have argued previously as my prior arguments have focused on facets that are either within-teacher or between-teacher. The CI facets are within-teacher facets (because they vary within-teachers between-days), but as this analyses shows have some between-teacher aspect. Co-construction-type effects are possible, but act within-teachers (under the assumption of a constant facet effect) because many teachers are observed across the full range of possible values of the facet (by definition). Differences between teachers are solely driven by the number of days teachers are observed at each level of the facet. This is in contrast to between-teacher facets (like average class prior achievement), which, incidentally,

differences in the frequency with which teachers engage in specific forms of instruction. In this case, we should prefer the unadjusted model because the adjusted model will statistically eliminate a source of true differences in teacher quality[47]. When the between-school component of a CI facet is significant (i.e. discussion on CLASS and PLATO), it implies that schools where teachers were observed teaching a level of a facet more often had higher (or lower) school-average observed teaching quality (e.g. schools where more observations of discussion lessons occurred had higher observed teaching quality). In this case, it is not clear which model to prefer because the difference could stem from school sources (e.g. a curriculum that promotes discussion lessons) or it could stem from teacher sources (e.g. teachers who choose to teach more discussion lessons choose to work at specific schools). If the school is the source, the adjusted model is preferred because the adjusted model removes the impact of the school context whereas the reverse is preferable if the teacher is the source. This shows how complex the problem of when to adjust for hidden facets can become. When facets are acting on multiple levels of nesting, the benefits and costs of using statistical models that adjust for the effects of facets must be balanced across the impact of adjusting at each nesting level.

Tables 5.8, 5.9, and 5.10 also show how the SO facets affected observed teaching quality within-teachers, between-teachers, and between-schools. Table 5.8 shows that prior

may vary within-teachers between-sections. For between-teacher facets, however, the within-teacher component cannot fully account for any co-construction effects because not enough teachers are sampled across the full range of possible values of the facet (note that 80% of the variance in the demographic composite is between schools so teachers cannot be observed across the full range of the variable since they are in a single school). Thus, co-construction can occur between-teachers because facet effects that drive co-construction may only occur between teachers.

[47] Note that I am assuming here that the effect of the facet on observed teaching quality is constant across teachers *or* the sampling of days for teachers is independent of the size of the effect of the facet for a given teacher (i.e. a teacher with a larger facet effect is not observed on that facet more often than a teacher with a smaller facet effect). I also assume all teachers within a school face the same set of contextual features after controlling for facets included in the model, such as grade, prior achievement, and demographic characteristics. If teachers in a school face unique contextual features (such as a special education teacher might) that lead them to engage in different types of instruction (which introduces heterogeneity within a level of the CI facet—a complication not addressed here), teacher sorting would not be the only possible between-teacher within-school effect that could cause this effect.

achievement was associated only with between-school differences in observed teaching quality for CLASS (ES=0.39 SD_T). Table 5.9 shows prior achievement was associated with between-teacher within-school (ES=0.50 SD_T) and between-school (ES=0.55 SD_T) differences in observed teaching quality on FFT. Table 5.10 shows no effect of prior achievement on PLATO at any level. The demographic composite, on the other hand, is related to between-teacher within-school differences in observed teaching quality only on CLASS (ES= -0.64 SD_T) and PLATO (ES= -0.89 SD_T), but not FFT, as can be seen in Tables 5.8, 5.10, and 5.9, respectively. Because none of the SO facets act within-teachers, the SO facets will not contribute to imprecise estimates of teacher quality ($\widehat{v_t}$), but they may contribute to bias in estimates of teacher quality ($\widehat{v_t}$). As I have discussed, if the facets affected observed teaching quality because of co-construction effects, then a model that adjusts for the hidden facets should be preferred because it "corrects" for effects on observed teaching quality not caused by the teacher (i.e. it equates the different contextual factors teachers face). If the facets affected observed teaching quality because of teacher sorting, then an unadjusted model should be preferred because the differences across the facets reflected true differences in teacher quality.

The analyses presented in this section focused on the level of nesting at which facets affected observed teaching quality. Throughout this thesis, I have argued that facets that act within-teachers contribute to imprecise estimates of teacher quality ($\widehat{v_t}$) and inflate estimates of the variance of teacher quality ($\widehat{var}(v_t)$) whereas between-teacher, within-school and between-school effects may or may not cause bias in estimates of teacher quality ($\widehat{v_t}$), depending on whether co-construction or teacher sorting is the cause of the effect and the model employed to estimate teacher quality. The analyses presented here show that the CI facets acted at all levels of nesting, which was a surprise given that the majority of the variance across these facets was within-teachers. The SO facets acted between-teachers with prior achievement acting between-schools on CLASS and FFT, prior achievement acting between-teachers, within-schools on FFT only, and the demographic composite acting between-teachers, within-schools on CLASS and PLATO. Thus, both CI and SO facets may contribute to biases in estimates of teacher quality. Further, there is no evidence in the UTQ data that can allow me to determine whether we should prefer models that do or do not adjust for these facet effects.

V.2.5. Change in Variance Components Across Models Since introducing the Base model, I have focused solely on the estimated effects of hidden facets on observed teaching quality. However, as hidden facets are added to the GTheory models, the size of the error facets changes. I focus in this section on a common Hierarchical Linear Modeling (Raudenbush & Bryk, 2001) approach that examines how the inclusion of explanatory variables (here, hidden facets) affects the variation of the model's random effects (here, the planned facets of measurement). Exploring changes to the random effects is important for developing a full understanding of how the hidden facets affected the measurement properties of observation instruments. For example, hidden facets that explain why days of instruction vary within-teachers will reduce the variance of the day facet while hidden facets that explain differences in teacher quality will reduce the variation of the teacher facet. By understanding how each category of hidden facet changed the variance of the planned facets of measurement, we learn how the hidden facets affect observed teaching quality. If, for example, the rater error facet variance is reduced to zero after controlling for a set of hidden facets, we would have identified a set of hidden facets that explains why some raters are more harsh or lenient than others. This can reveal a lot about the nature of the planned facets of measurement.

Table 5.11 shows how the variances of the planned facets of measurement change across models for CLASS. Tables 5.12 and 5.13 show the same for FFT and PLATO, respectively. The tables consist of three sets of comparisons. Each comparison shows the difference in the variances of the facet across two models. The first three columns compare the Base model (Base) to the System Design (SD) model. The next set of three columns compares the Base model to the Curriculum and Instruction (CI) model. The last set of columns compares the Base model to the School Organization (SO) model. Within each set of columns, the left column presents the variance of the indicated error facet from the Base model (and so it is the same for each set of columns). The middle column presents the variance of indicated error facet from the comparison model. The right column shows the percentage of change (i.e. (x-y)/x) across the two models. Additionally, Tables A.1, A.2, and A.3 in Appendix A show variances of the error facets with 95% confidence intervals across each model, allowing for a sense of whether changes are statistically significant[48].

[48] While the percentage change in facet size across models is often large, the differences are generally smaller than the uncertainty in estimates. That is, changes are non-significant. The only exception is for the teacher facet of the SO model for all instruments and the teacher facet of the CI model for PLATO. This is caused by the large relative uncertainty in estimates, which results in very large percentage changes in variance components being necessary in order to get significant changes. Thus, differences in the variance components across models should not be over-interpreted. The differences are large enough to be meaningful, however, and presenting uncertainty in variance components is not standard practice so I still briefly discuss the implications of the differences in this section.

Table 5.11: Change in the Variance of the Error Facets across the CLASS Models

	SD Model Comparison			CI Model Comparison			SO Model Comparison		
Facet	Base	SD	Perc Change	Base	CI	Perc Change	Base	SO	Perc Change
Teacher ($var(v_t)$)	0.076	0.066	14%	0.076	0.06	21%	0.076	0.031	59.2%
Day ($var(v_{d:s:t})$)	0.013	0.007	44%	0.013	0.007	47%	0.013	0.005	60.8%
Occasion ($var(v_{o:d:s:t})$)	0.053	0.052	3%	0.053	0.052	3%	0.053	0.052	3.0%
Rater ($var(v_r)$)	0.04	0.021	46%	0.04	0.022	45%	0.040	0.022	44.8%
Rater-by-Teacher ($var(v_{rt})$)	0	0.012	0%	0	0.013	0%	0.000	0.014	0%
Rater-by-Day ($var(v_{r(d:s:t)})$)	0.141	0.116	18%	0.141	0.114	19%	0.141	0.112	20.3%
Item-by-Rater ($var(v_{ir})$)	0.225	0.225	0%	0.225	0.225	0%	0.225	0.225	0.0%
Item-by-Teacher ($var(v_{it})$)	0.029	0.029	-1%	0.029	0.029	-1%	0.029	0.029	-0.8%
Item-by-Day ($var(v_{i(d:s:t)})$)	0.128	0.129	-1%	0.128	0.129	-1%	0.128	0.129	-0.7%
Item-by-Occasion ($var(v_{i(o:d:s:t)})$)	0	0	100%	0	0	100%	0.000	0.000	80.1%
Residual ($var(\epsilon_{ir(o:d:s:t)})$)	0.381	0.377	1%	0.381	0.377	1%	0.381	0.377	1.0%

Note. Each table consists of three sets of comparisons. Each comparison shows the difference in error facet variances across two models. The first three columns compare the base model (Base) to the System Design (SD) model. The next set of three columns compares the Base model to the Curriculum and Instruction (CI) model. The last set of columns compares the Base model to the School Organization (SO) model. Within each set, the left column presents the variance of the indicated error facet from the Base model. The middle column presents the variance of indicated error facet from the comparison model. The right column shows the percentage of change (i.e. (x-y)/x) across the two models.

Table 5.11 shows the results for CLASS. Table 5.11 shows that the variance of the teacher facet (i.e. $var(v_t^{model})$) decreased by 14% from the Base to the SD model, by 21% from the Base to the CI model, and by 59% from the Base to the SO model. This implies that 14% of the variation of the teacher facet in the Base models was attributable to when teachers were observed and how scoring was organized (i.e. the SD facets). A further 7% of the variation in the teacher facet in the Base model was attributable to differences in the content domain being taught and the interaction structures being used when teachers were observed. This result confirmed my prediction of an inflated variance estimate in the Base model stemming from the effects of within-teacher hidden facet effects (assuming the hidden facets effects capture error or within-teacher effects). Finally, a further 38% the variation in the teacher facet in the Base model was explainable by between-teacher differences in the students being taught and the grade being taught (i.e. SO facets). If co-construction explains the SO facet effects, this variance is error and should not affect teacher score estimates (and it is not error if teacher sorting explains the SO facet effects). Thus, these results suggest that observation instruments may be incorporating a lot of non-teacher variation in their estimates of teacher quality and the choice of which facets to adjust for will have a major impact on how well teacher quality is being measured. As I show below, the net effect of these results was to reduce the reliability of teacher scores as more controls for hidden facets were added to the model.

Table 5.11 also shows that other planned facets of measurement change across the different models. The variance of the day facet, the rater facet, and the rater-by-day facet changed a lot across the models[49]. Forty-four percent of the day variance in the Base model was explained by the SD facets while the CI facets explained a further 3% and the SO facets explained a further 14% of the variance of the day facet in the Base model. Thus, differences in the time of year observed, the day of the week observed, rater drift, and the scoring mode explained almost half of the variance in observed teaching quality between-days within-teachers, while student characteristics and grade taught (i.e. the SO facets) explained an additional portion of the variance of the day facet in the Base model. Overall, across all categories of facets, Table 5.11 shows that 61% of the variance in the day facet in the Base model was explained. Similarly, the SD model explained 46% of the variance of the rater facet, indicating that rater drift, scoring mode, and when observations occurred played a significant role in determining differences in rater severity. The CI model and SO model did not further reduce the variance of the rater facet. Table 5.11 also shows that the variance of the rater-by-day facet was reduced by 18% by the SD facets, showing the explanatory value of these facets for another source of rater error. Thus, much of the rater error in CLASS is the result of systematic effects related to the SD facets.

[49] I do not consider the item-by-occasion facet because the Base model shows the variance across this facet is almost zero.

Table 5.12 shows the same results, but for FFT. Looking at table 5.12, we see that on FFT, 13% of the teacher facet variance in the Base model was explained by the SD model, 21% was explained by the CI model, and 61% was explained by the SO model. This is a remarkably similar to the results from CLASS shown in Table 5.11. Table 5.12 also shows that the day and rater-by-day facets changed across the different models. Of the variance in the day facet in the Base model, 3% was explained by the SD model, 7% was explained by the CI model, and 11% was explained by the SO model. Thus, the hidden facets explained significantly less variance across the day facet on FFT than on CLASS, showing that the variation within-teachers between-days in FFT scores was due to different sources than the same type of variation in CLASS scores. Last, of the variance in the rater-by-day facet on FFT, 12% was explained by the SD model while the CI and SO models did not further explain the variation across the rater-by-day facet. Similar to CLASS, then, the SD facets (such as scoring mode or rater drift) explained a fairly large fraction of the variation across the rater-by-day facet.

Table 5.12: Change in the Variance of the Error Facets across the FFT Models

	SD Model Comparison			CI Model Comparison			SO Model Comparison		
Facet	Base	SD	Perc Change	Base	CI	Perc Change	Base	SO	Perc Change
Teacher ($var(\upsilon_t)$)	0.029	0.026	12.6%	0.029	0.023	21.1%	0.029	0.011	61.3%
Day ($var(\upsilon_{d:s:t})$)	0.008	0.008	3.2%	0.008	0.008	6.6%	0.008	0.007	10.9%
Rater ($var(\upsilon_r)$)	0.011	0.011	5.1%	0.011	0.010	10.0%	0.011	0.011	4.9%
Rater-by-Teacher ($var(\upsilon_{rt})$)	0.005	0.005	-1.1%	0.005	0.005	-5.3%	0.005	0.005	-4.6%
Rater-by-Day ($var(\upsilon_{r(d:s:t)})$)	0.044	0.038	12.3%	0.044	0.038	12.8%	0.044	0.038	13.4%
Item-by-Rater ($var(\upsilon_{ir})$)	0.011	0.011	-0.1%	0.011	0.011	-0.1%	0.011	0.011	-0.1%
Item-by-Teacher ($var(\upsilon_{it})$)	0.008	0.008	1.0%	0.008	0.008	1.0%	0.008	0.008	1.0%
Item-by-Day ($var(\upsilon_{i(d:s:t)})$)	0.017	0.017	-1.2%	0.017	0.017	-1.2%	0.017	0.017	-1.2%
Residual ($var(\epsilon_{ir(o:d:s:t)})$)	0.140	0.140	0.2%	0.140	0.140	0.2%	0.140	0.140	0.2%

Note. Each table consists of three sets of comparisons. Each comparison shows the difference in error facet variances across two models. The first three columns compare the base model (Base) to the System Design (SD) model. The next set of three columns compares the Base model to the Curriculum and Instruction (CI) model. The last set of columns compares the Base model to the School Organization (SO) model. Within each set, the left column presents the variance of the indicated error facet from the Base model. The middle column presents the variance of indicated error facet from the comparison model. The right column shows the percentage of change (i.e. (x-y)/x) across the two models.

Table 5.13 shows the same results, but for PLATO. As Table 5.13 shows, compared to the Base model, the variance of the teacher facet was 16% lower on the SD model, 38%

lower on the CI model, and 57% lower on the SO model. The amount of reduction in the variance across teachers in the SD and SO models matched the results from CLASS and FFT, but the CI model explained more of the variation across teachers in observed teaching quality on PLATO than on the other two instruments. This was due to the larger effect of the CI facets on PLATO scores, which is turn was likely either driven by the same rater providing PLATO scores and PLATO log responses used to create the CI facets (i.e. correlated rater error) or could reflect PLATO score's greater sensitivity to specific instructional practices that shift across levels of the CI facets. The reader will notice from Table 5.13 that the percent change in the variance components across models for PLATO scores were far more variable than for CLASS and FFT. Much of this can be explained by the variances of many facets being non-significantly larger than zero in the Base model, which suggests this variation across models may be sampling error shifting estimates as the models change. Table A.3 in Appendix A shows the value of the variances with 95% confidence intervals for PLATO across the four models. As Table A.3 shows, the estimated variance of the day facet, rater facet, and rater-by-teacher facet had confidence intervals that included zero, so I do not interpret the effects of these facets in Table 5.13 (as they likely represent sampling error). The variance of the item-by-occasion facet decreased by 56% when moving from the Base model to the SD model and remained the same size in the CI and SO models. This likely reflected the impact of the item by occasion order interaction effects included in the SD model.

Table 5.13: Change in the Variance of the Error Facets across the PLATO Models

	SD Model Comparison			CI Model Comparison			SO Model Comparison		
Facet	Base	SD	Perc Change	Base	CI	Perc Change	Base	SO	Perc Change
Teacher ($var(v_t)$)	0.012	0.010	16%	0.012	0.007	38%	0.012	0.005	57%
Day ($var(v_{d:s:t})$)	0.003	0.004	-8%	0.003	0.000	100%	0.003	0.000	100%
Occasion ($var(v_{o:d:s:t})$)	0.017	0.016	4%	0.017	0.016	3%	0.017	0.016	3%
Rater ($var(v_r)$)	0.002	0.000	100%	0.002	0.000	90%	0.002	0.000	85%
Rater-by-Teacher ($var(v_{rt})$)	0.000	0.000	100%	0.000	0.000	100%	0.000	0.000	43%
Rater-by-Day ($var(v_{r(d:s:t)})$)	0.020	0.019	6%	0.02	0.019	7%	0.02	0.018	8%
Item-by-Rater ($var(v_{ir})$)	0.022	0.021	1%	0.022	0.021	1%	0.022	0.021	1%
Item-by-Teacher ($var(v_{it})$)	0.012	0.013	-4%	0.012	0.013	-4%	0.012	0.013	-4%
Item-by-Day ($var(v_{i(d:s:t)})$)	0.067	0.069	-3%	0.067	0.069	-3%	0.067	0.069	-3%
Item-by-Occasion ($var(v_{i(o:d:s:t)})$)	0.012	0.005	56%	0.012	0.005	56%	0.012	0.005	56%
Residual ($var(\epsilon_{ir(o:d:s:t)})$)	0.246	0.245	0%	0.246	0.245	0%	0.246	0.245	0%

Note. Each table consists of three sets of comparisons. Each comparison shows the difference in error facet variances across two models. The first three columns compare the base model (Base) to the System Design (SD) model. The next set of three columns compares the Base model to the Curriculum and Instruction (CI) model. The last set of columns compares the Base model to the School Organization (SO) model. Within each set, the left column presents the variance of the indicated error facet from the Base model. The middle column presents the variance of indicated error facet from the comparison model. The right column shows the percentage of change (i.e. (x-y)/x) across the two models.

The results presented in this section showed that the hidden facets included in the statistical model will determine, to some extent, the degree to which observed teaching quality scores vary across the planned measurement facets. As I had predicted, the within-teacher SD facets caused the Base model to have arguably inflated estimates of the variance across teachers. Across the instruments, the variance of the teacher facet was about 15% lower in the SD model than in the Base model, implying that, assuming the SD facets do not contribute to teacher quality, 15% of the variance in the teacher score estimates from the Base model was actually sampling error coming from the SD facets. Further, the results showed that a further 8% of the variance in the teacher scores from the Base model was explained by the CI facets on CLASS and FFT, while a further 22% of the variance in the teacher scores from the Base model on PLATO was explained by the CI facets, which also may be sampling error if the frequency of observing a teacher at a level of a hidden facet does not predict the frequency of that teacher teaching at that level of the hidden facet *or* the frequency with which teachers engage in instruction at different levels of the hidden facet is not considered an aspect of teacher quality. If we further assume that co-construction caused the SO facet effects, then a total of about 60% of the variance in estimated teacher scores from the Base model was due to error stemming from the SD, CI, or SO facets. These effects were remarkably consistent across instruments, though PLATO had a larger reduction in teacher facet variance due to the CI facets. Beyond the teacher facet, this consistency breaks down. Controlling for the SD facets explained a great deal of the rater error on CLASS, but not FFT or PLATO. Similarly, controlling for the SD facets explained a great deal of the rater-by-day error facet on CLASS and FFT, but much less for PLATO. These both suggest that the source of rater error varies across instruments. Importantly, the analyses shown here suggest that, assuming the SO model is correct, the variation in observed teaching quality that was attributable to teacher differences was drastically over-estimated, to the extent that, as I show below, the SO model seems to have little ability to reliably distinguish between teachers possessing different levels of teacher quality.

V.3. Impact of Hidden Facets on Estimated Teacher Quality

To this point, I have focused on the effect of hidden facets on observed teaching quality and changes to the estimated variances across the statistical models. In this section, I begin to focus on research question 2c: How much does adjusting for the contextual features of measurement (i.e. hidden facets) change estimated teacher quality scores and estimated score reliability? To explore this question, I extract estimates of teacher quality (e.g. υ_t^{base}, υ_t^{SD}, ...) from the models described earlier. I then make comparisons of these scores to each other, the simple teacher average score, and occasionally other teacher score estimates. Further, I explore how the reliability of these scores changes across models. The goal of this section is to explore whether the hidden facets lead to meaningfully different score estimates. I start by looking at the simple correlation of teacher score estimates across models. I then move to examine how a teacher's rank in the teaching quality distribution changes across models, which gives a view of the effect of adjusting for hidden facets that captures the impact on individual teachers. Last, I look at how estimated decision study reliability changes across models.

V.3.1. Teacher Score Correlations across Models A common way to explore how much estimates of teacher quality change across different models is to examine the simple correlation between scores derived from these different models. This very broadly addresses the question of whether adjusting for hidden facets makes a practical difference when it comes to teacher score estimates. Tables 5.14, 5.15, and 5.16 show the correlation of teacher scores across the models for CLASS, FFT, and PLATO, respectively. Beyond using teacher scores estimated from the Base model, SD model, CI model, and SO model, which have been extensively discussed, I create estimates of teacher quality using three additional approaches. First, I simply average observed scores up to the teacher-level (Ave). Second, I averaged scores after removing rater main effects (Rater). Last, I include teacher score estimates from a model that is identical to the Base model, but includes a school random effect so that teacher quality captures only within-school differences in quality (BaseW)[50].

As Tables 5.14, 5.15, and 5.16 show, the correlations of scores across all models was very high. Correlations mostly remained above 0.95 for scores formed from averaging the observed scores (Ave), averaging the observed scores after removing the rater main effects (Rater), the Base model, the SD model, and the CI model. This implies that, if teacher score estimates are the only concern, the simple average give the same result as more complicated statistical models. However, this relies on the random sampling of days in the UTQ data so likely does not apply to data where sampling is not carefully conducted. This was surprising

[50] The within-school model is not ideal because teachers in UTQ are a non-random sample of the teachers within their school. Thus, the deviation is not from the true school average, but an estimate of that average from a set of non-randomly selected teachers. This is problematic if teachers in different schools were selected into UTQ under different mechanisms. It is still, however, interesting to explore these effects.

given the large effects of the hidden facets on observed teaching quality in the SD and CI models discussed earlier. Apparently, the random sampling across days in the UTQ sampling design helped average out the effects of the SD and CI facets, thereby minimizing their impact on estimated teacher quality. In addition, the fact that there was no correlation between teacher quality and likelihood of being observed on any SD or CI facet (due to the random sampling of days) also helps minimize the effects of these facets on teacher quality estimates. When sampling is less well-controlled, this averaging out of the effects of SD and CI facets should not be expected and the difference in estimated teacher scores across models will likely be much larger.

Importantly, however, the correlations of teacher score estimates from the SO model to the Base model were lower (0.76-0.82). Finally, the correlation of estimated scores from the within-school Base model (BaseW in the tables) and the SO model are higher than for any other model while the BaseW model correlates with other models more strongly than the SO model. This places BaseW scores between the scores from the Base model and the SO model, a sort of compromise between models not adjusting for SO facets and those explicitly adjusting for these facets. This within-school Base model (BaseW) allows extrapolation of scores across schools under the assumption that teachers who received the highest score in their current school will receive the highest score in all schools (i.e. schools have a mean impact on observed teaching quality that is constant across all teachers) while the Base model assumes teacher sorting effects and the SO model assumes co-construction effects. This is an alternative way of supporting the extrapolation of scores across schools that has been used in VA scores due to concerns about the difficulty of comparing teachers across schools (Reardon & Raudenbush, 2009), concerns which also apply to classroom observation scores.

Table 5.14: Correlation of Teacher CLASS Scores Estimates across Models

	Ave	Rater	Base	SD	CI	SO	BaseW
Ave	1	0.98	0.97	0.94	0.93	0.76	0.87
Rater	0.98	1	0.99	0.96	0.95	0.76	0.89
Base	0.97	0.99	1	0.97	0.96	0.76	0.89
SD	0.94	0.96	0.97	1	0.99	0.82	0.90
CI	0.93	0.95	0.96	0.99	1	0.83	0.90
SO	0.76	0.76	0.76	0.82	0.83	1	0.84
BaseW	0.87	0.89	0.89	0.90	0.90	0.84	1

Note. Ave is the observed teaching quality score averaged to the teacher-level. Rater is the observed teaching quality score averaged to the teacher-level with rater main effects removed. Base is teacher score estimate from the Base Model. SD is teacher score estimate from the System Design Model. CI is teacher score estimate from the Curriculum and Instruction Model. SO is teacher score estimate from the School Organization Model. BaseW is the teacher score estimates from the Base model, but centered within schools.

Table 5.15: Correlation of Teacher FFT Scores Estimates across Models

	Ave	Rater	Base	SD	CI	SO	BaseW
Ave	1	0.98	0.98	0.95	0.94	0.76	0.88
Rater	0.98	1	0.99	0.97	0.96	0.77	0.90
Base	0.98	0.99	1	0.98	0.96	0.77	0.91
SD	0.95	0.97	0.98	1	0.99	0.81	0.92
CI	0.94	0.96	0.96	0.99	1	0.84	0.92
SO	0.76	0.77	0.77	0.81	0.84	1	0.85
BaseW	0.88	0.90	0.91	0.92	0.92	0.85	1

Note. Ave is the observed teaching quality score averaged to the teacher-level. Rater is the observed teaching quality score averaged to the teacher-level with rater main effects removed. Base is teacher score estimate from the Base Model. SD is teacher score estimate from the System Design Model. CI is teacher score estimate from the Curriculum and Instruction Model. SO is teacher score estimate from the School Organization Model. BaseW is the teacher score estimates from the Base model, but centered within schools.

Table 5.16: Correlation of Teacher PLATO Scores Estimates across Models

	Ave	Rater	Base	SD	CI	SO	BaseW
Ave	1	0.99	0.98	0.95	0.91	0.81	0.92
Rater	0.99	1	0.99	0.97	0.93	0.83	0.94
Base	0.98	0.99	1	0.98	0.93	0.82	0.94
SD	0.95	0.97	0.98	1	0.95	0.86	0.95
CI	0.91	0.93	0.93	0.95	1	0.92	0.91
SO	0.81	0.83	0.82	0.86	0.92	1	0.87
BaseW	0.92	0.94	0.94	0.95	0.91	0.87	1

Note. Ave is the observed teaching quality score averaged to the teacher-level. Rater is the observed teaching quality score averaged to the teacher-level with rater main effects removed. Base is teacher score estimate from the Base Model. SD is teacher score estimate from the System Design Model. CI is teacher score estimate from the Curriculum and Instruction Model. SO is teacher score estimate from the School Organization Model. BaseW is the teacher score estimates from the Base model, but centered within schools.

In summary, then, Tables 5.14, 5.15, and 5.16 show that adjusting for the SD and CI hidden facets had minimal effects on estimates of teacher quality scores (as compared to estimates from the Base model or simple mean scores) while teacher quality estimates from the SO model had slightly lower correlations with teacher quality estimates from the Base model. On the basis of these findings, adjusting for SD and CI facets (or even using a model beyond simply taking means) might not be worth the effort when the goal is only to estimate teacher quality scores, *although* the reader should take note that the results reported here are probably the result of the UTQ sampling design, which selected days of observation more or less at random. If, on the other hand, an observation protocol selected days of observation in a way that was correlated to teaching quality (as might be the case of teachers wanted to be observed teaching their best curriculum), the correlations between adjusted and unadjusted scores might not show the patterns found in Tables 5.14 – 5.16. In addition, the findings suggest that a failure to adjust for SO facets (related to student composition) could be more consequential, as correlations of simple models and models adjusted for SO facets are only correlated in the range of .75 - .85. Though, as we have seen, any adjustment of SO facets involves the question of what assumptions are appropriate to generalize observed teaching quality across the SO facets, a question I return to in the discussion.

V.3.2. Difference in Rank Scores across Models The correlation of teacher score estimates from different models just discussed provides a broad view of how much adjusting for hidden facets affects estimates of teacher quality (i.e. teacher scores; $\widehat{\upsilon_t}$). However, in many practical settings, teachers will face individual consequences for their scores, so in this section I change from looking at the impact of model-to-model variation in scores to how specific teachers will shift their location in the distribution of teacher scores as estimation models change. I do this by examining how much teacher score estimates change ranks in the distribution of teacher quality as estimating models change. For example, a teacher's estimated score might be in the 54th percentile of the distribution in the Base model, but move to the 68th percentile when using the SD model. This switch would thus move the teacher's score by 14 ranks. By calculating this rank shift across all teachers, we can ask how many ranks did the 1% of teachers who showed the greatest change experience. This provides information about how much the specific model used to estimate teacher quality scores affects the value of those scores for individual teachers (say for the 1% of teachers who experienced the greatest change). This is the same information conveyed by the correlations above, but the results are more interpretable as the implications of the statistical model for individual teachers.

Table 5.17 shows the results of this analysis[51]. The first two columns specify the instrument and the target model that is being switched to. I look only at how much switching from the Base model to the target model affects a teacher's score's rank. The percentages show quantiles of the distribution of the difference in teacher's score's rank across models. Consider the top row of Table 5.17. This row shows the effect of moving from the Base model to the SD model for the teacher's rank on CLASS. Ninety-percent of teachers will shift their rank by 1 percentile point or more; 75% by 2 percentile points or more; and 10% will shift by 14 percentile points or more. Thus, about 23 teaches had their scores move over 1/10 of the distribution when moving from the Base model to the SD model. The one percent of teachers who experienced the largest change between the Base model and the SD model shift 20 percentile points in CLASS, 14 in PLATO, and 18 in FFT. The changes are only slightly larger for the CI model, except for PLATO. Table 5.17 shows that moving to the SO model has larger implications for teachers. One-percent of teachers will move almost 50 percentile points across the distribution while one-quarter of teachers will move 20-26 percentile points or more. The analyses presented here, then, show that while controlling for hidden facets on the estimated teacher quality scores was relatively minor sample-wide (i.e. correlations were high across models), the decision of what hidden facets to adjust for can be very consequential for individual teachers.

[51] Note there are 228 teachers so 1% of teachers is 2.28 teachers.

Table 5.17: Percentile Shift in the Rank of Teacher's Score Estimates across Models compared to the Base Model

Instrument	Target Model	90%	75%	50%	25%	10%	1%
CLASS	System Design	1	2	5	9	14	20
PLATO	System Design	0	2	4	7	11	14
FFT	System Design	1	2	4	8	11	18
CLASS	Curriculum/Instruction	1	2	6	11	17	23
PLATO	Curriculum/Instruction	1	3	7	14	20	33
FFT	Curriculum/Instruction	1	2	5	10	14	23
CLASS	School Organization	2	5	14	26	37	50
PLATO	School Organization	1	4	11	20	31	49
FFT	School Organization	2	6	13	25	37	53

Note. Table shows the change in rank of the teacher quality distribution that a teacher's estimated score will experience when shifting from the Base model to the target model. For each row, a distribution of how many ranks teacher's scores change across models is formed. From the third row on, the cells display the quantiles of this distribution. The third column shows the minimum shift in rank that 90% of teachers will experience when teacher score estimates are estimated from the indicated model instead of the Base model. Fourth column shows the minimum shift in rank 75% of teachers will experience, and so on. * p<0.05; ** p<0.01; *** p<0.001.

V.3.3. Change in Reliability across Models In this section, I examine how estimates of score reliabilities change across models. In a previous section, I showed that estimates of the variance in teacher scores decreased as more controls were added to adjust for hidden facets. This has implications for score reliability, suggesting the reliability will decrease as models add controls for hidden facets. The reader will recall that this is what I predicted in earlier chapters, where I argued that the presence of hidden facets increases the sampling error of observation scores, thereby providing artificially high estimates of the reliability of teacher scores derived from models that do not include the facets. In this section, I explore this idea and in doing so, show what is one of the largest effects of controlling for hidden facets, namely changes in estimated score reliability.

I present the results of these analyses in Figures 5.2, 5.3, and 5.4. These figures graphically represent the results from decision studies I conducted for each instrument and show the implications for reliability estimates of using different estimation models across different combinations of days observed and raters per day used in an observation system. I focus on these two variations in observation System Design because these are two main design features of most observational systems since most studies assume that the number of items is fixed by the choice of observation instrument. Note that Figures 5.2 – 5.4 do not include results for the SD model. That is because the results for that model almost exactly duplicate results for the CI model. In addition to the figures discussed next, I also will present the same data on reliability where there is only one choice for the number of raters per day. This allows me to show the 95% confidence intervals for the reliability estimates very clearly. Appendix C presents these results in a tabular format.

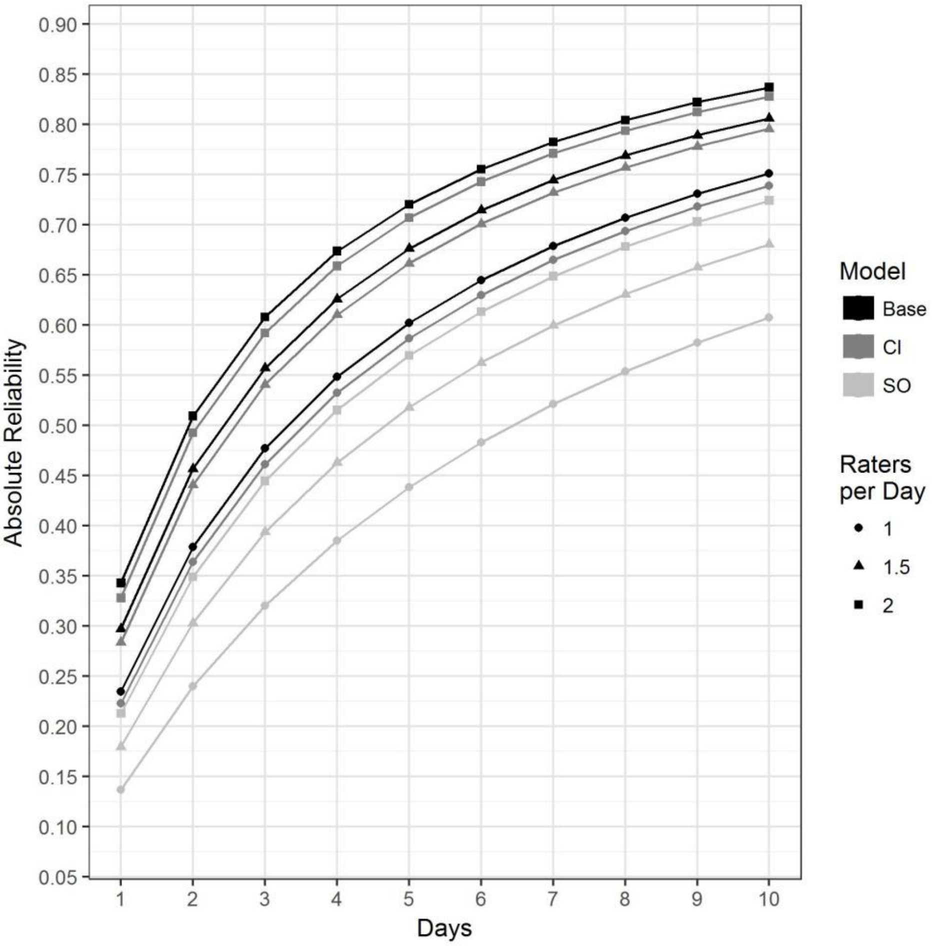

Figure 5.2: Estimated D-Study Reliability of the Teacher Score Estimate for CLASS across days, raters, and models

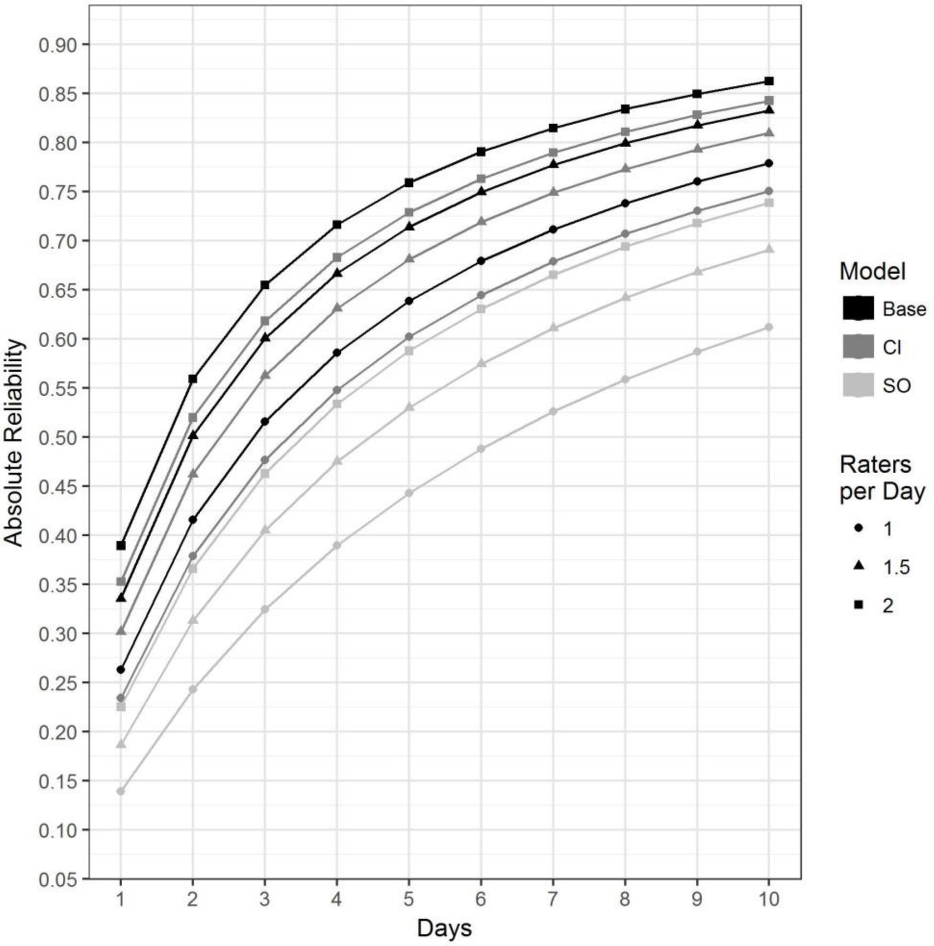

Figure 5.3: Estimated D-Study Reliability of the Teacher Score Estimate for FFT across days, raters, and models

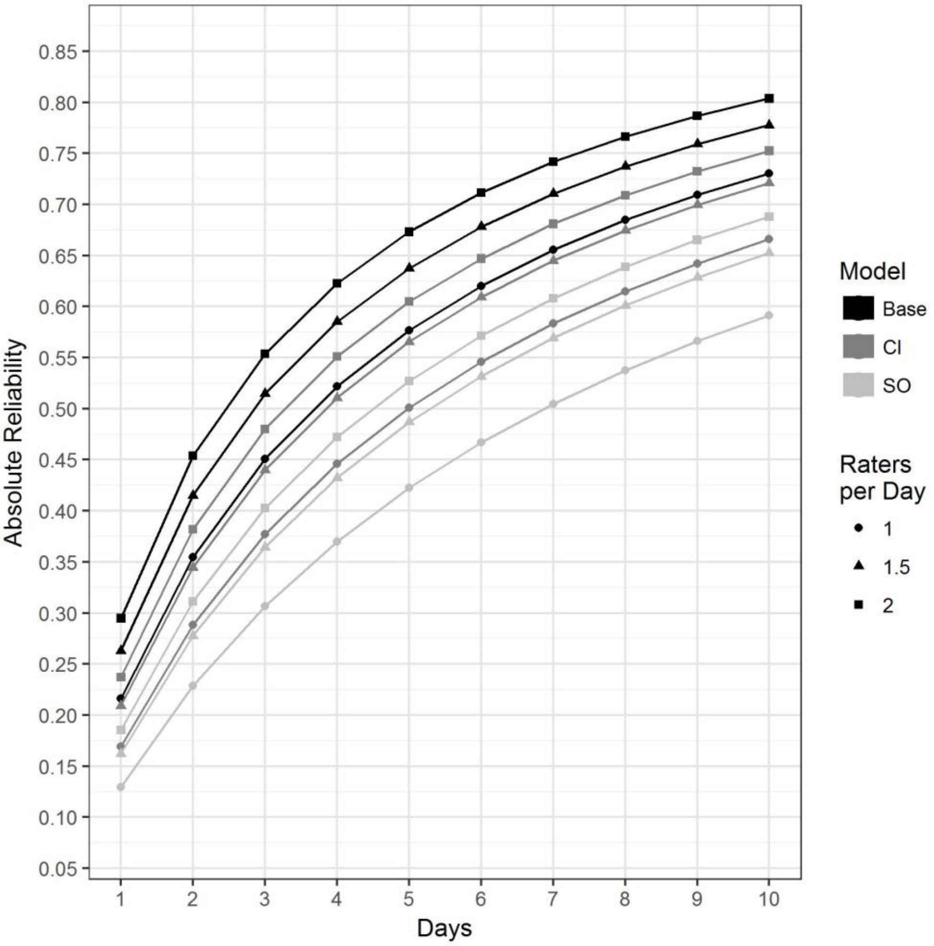

Figure 5.4: Estimated D-Study Reliability of the Teacher Score Estimate for PLATO across days, raters, and models

Figures 5.2, 5.3, and 5.4 show the results for CLASS, FFT, and PLATO respectively. It is interesting to note how similar the results were across instruments. I focus my discussion on an observation system that includes four days of teacher observation with a different rater each day. As the figures show, as more adjustments are made for hidden facets, estimated score reliability falls. This is because the adjusted models control away some part of the variation that the Base model attributes to true differences across teachers, reducing the variance of the "true score". The decrease was modest for the CI model, with estimated reliabilities moving from 0.55 to 0.53 on CLASS, 0.59 to 0.55 on FFT, and 0.52 to 0.45 on PLATO. Finally, the SO model had even lower reliabilities, just 0.39 for CLASS and FFT and 0.37 for PLATO.

Note that the reliabilities reported here are somewhat lower than those calculated from MET data, especially for PLATO, which had a reliability of 0.67 when observing 4 days in the MET study (Kane et al., 2012) and has a reliability of 0.53 for the equivalent design in the UTQ data. This difference between the two studies that arose because the teacher facet contributed relatively little variation to PLATO scores in the UTQ data as compared to in the MET data. Additionally, while the differences between score reliability in the Base model and the CI model were quite modest, they were large enough for FFT and PLATO to suggest that an additional day of observation is needed to maintain the same score reliability (i.e. they suggested moving from 4 to 5 days to maintain the score reliability). Observing an additional day entails a large financial cost so even this modest decrease is important. In fact, for the SO model, neither adding a second rater to score each day of instruction nor adding an additional day of observation will bring the reliability estimates up to that of the Base model.

Looking at Figures 5.5, 5.6, and 5.7, we can see the uncertainty of these reliability estimates with the 95% confidence interval of the estimates. The 95% confidence intervals span almost 0.2 points, which is large enough to generate considerable uncertainty in how many days of instruction and raters scoring each day are necessary for reliable teacher estimates. In fact, after observing for 3-4 days, adding an additional day of instruction does not make a statistically significant improvement in reliability estimates. Additionally, the CI model does not produce reliability estimates that are significantly lower than the Base model, though estimates of score reliability from the SO model are significantly lower than those from the Base model. Overall, then, the uncertainty in estimates makes it difficult to make definitive decisions on the number of days of instruction that should be observed, the number of raters scoring each day, and even the effect of moving from the Base model to the CI model.

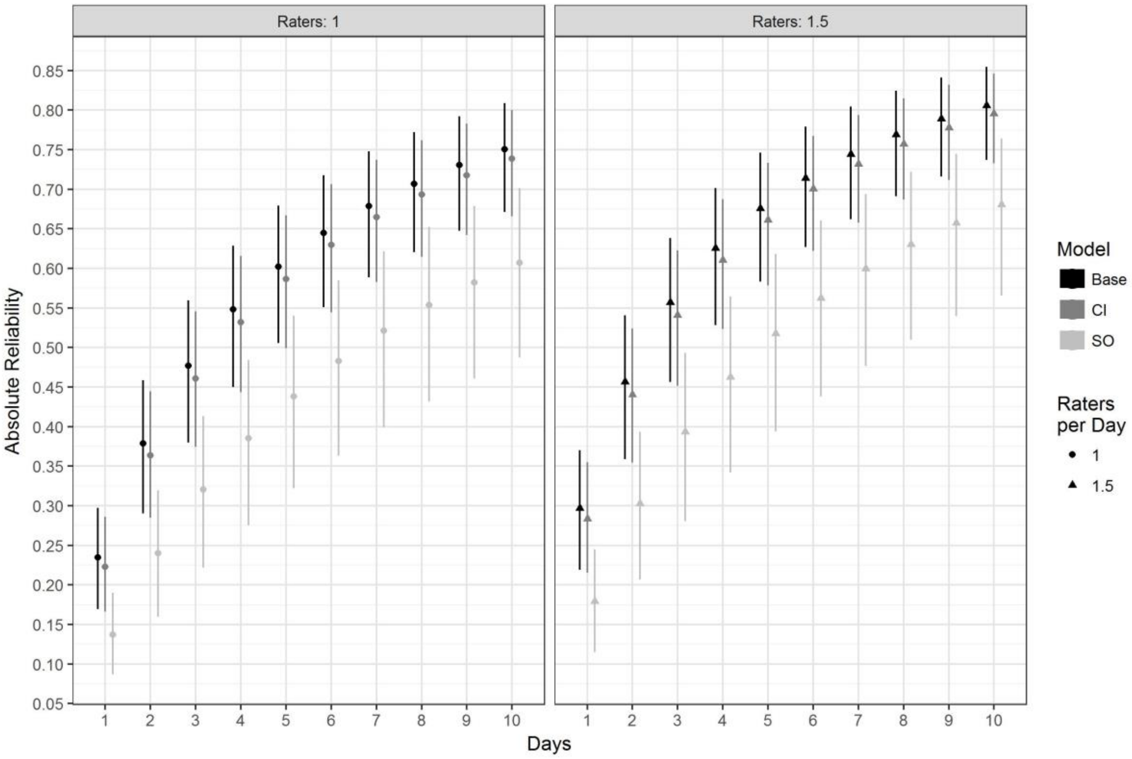

Figure 5.5: Estimated D-Study Reliability of the Teacher Score Estimate for CLASS with 95% CI across days, raters, and models

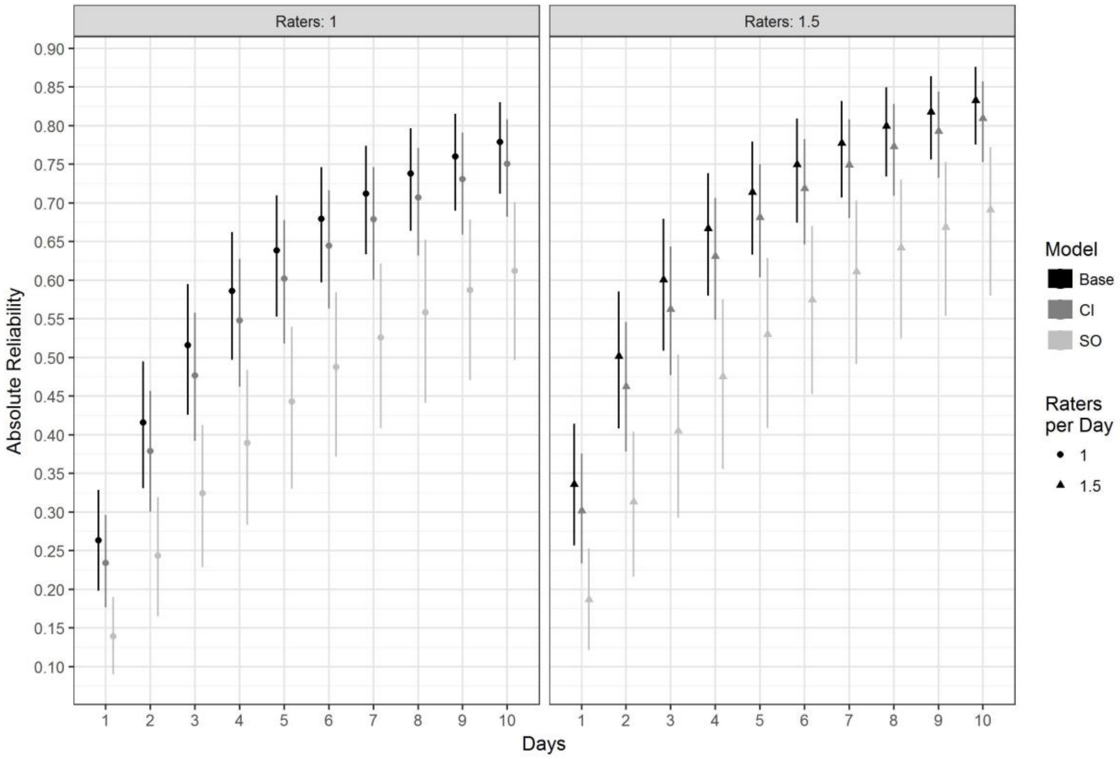

Figure 5.6: Estimated D-Study Reliability of the Teacher Score Estimate for FFT with 95% CI across days, raters, and models

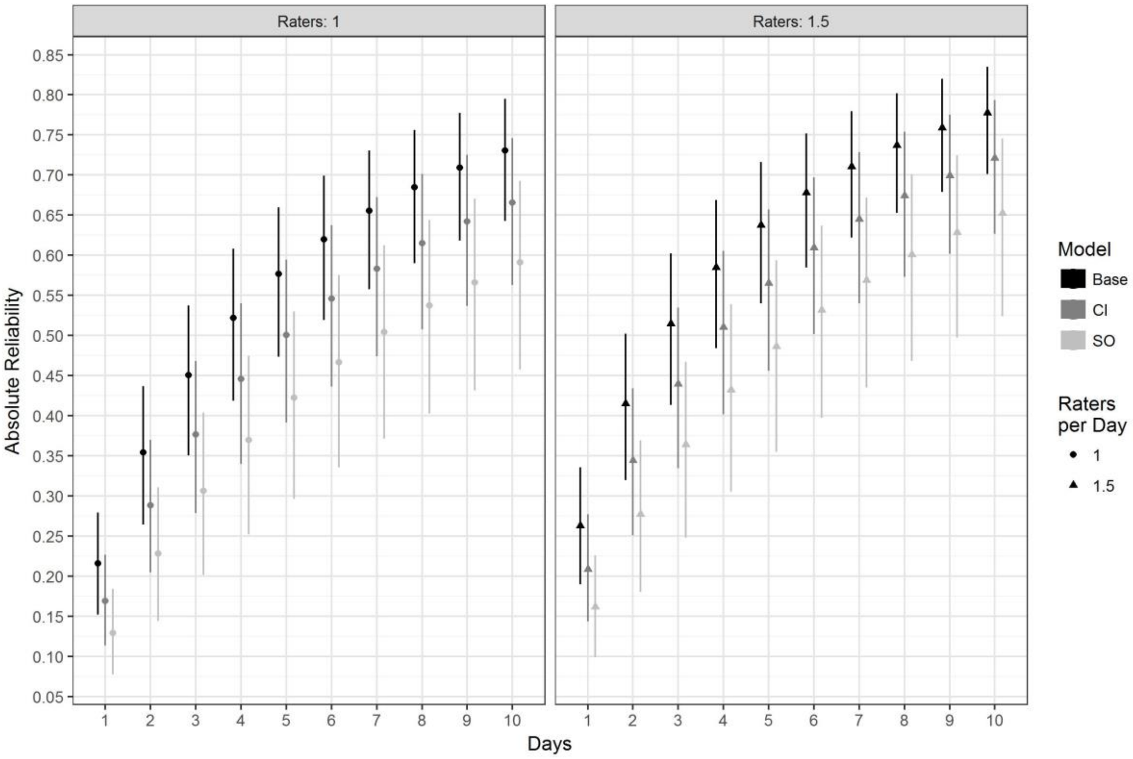

Figure 5.7: Estimated D-Study Reliability of the Teacher Score Estimate for PLATO with 95% CI across days, raters, and models

There is one major take-away from these reliability analyses. The reliability of scores estimated from the Base model appears to be positively biased. This bias is small relative to the uncertainty in the estimated reliability for the SD and CI facets, but large enough to change decisions about the design of an observational system. The effect is much larger for the SO facets, but, as I have discussed, whether the SO facets show a bias in the estimated reliability of the teacher quality scores is not straightforward. If we treat the adjustment for the SO facets as a proper correction for different circumstances of teaching (i.e. co-construction), the estimated reliabilities from the Base model (and in the current literature), drastically over-estimate score reliability—to the point where, under feasible System Design parameters, scores will never be adequately reliable. On the other hand, if one assumes that the effects of the SO facets on observed teaching quality is a result of teacher sorting across schools, the estimated reliabilities from the Base model are only slightly inflated (from effects of the SD and CI facets). This highlights the importance of determining whether SO facets capture a teacher sorting effect or co-construction-like effect.

V.4. Validity of Observation Scores

In this section, I address the third research question (RQ 3), which concerns the validity of the inference that observation scores capture teacher quality. Specifically, I will examine whether adjusting teacher quality scores for the effects of the SD, CI, and SO facets affects the validity of adjusted scores, where the validity data come from examining the (partial) correlation between the relevant teacher quality score and a teacher's value-added (VA) score as calculated by UTQ researchers. It is, of course, of relatively little interest how much observation scores correlate with VA scores, but instead the goal is to understand to what extent estimates of teacher quality represent the intended construct of teacher quality (i.e. the validity of scores). The correlation with VA scores provides a single (of many possible) view into how well teacher quality estimates capture the construct of teacher quality. In general, the higher the correlation of an estimate of teacher quality with VA scores, the more evidence exists for the validity of the observation scores, though, as I discuss below, this over-simplifies reality. Note that this is a narrow way of conceptualizing the validity of observational scores. Teacher quality is a broad construct with many different components, and so it is not necessarily the case that the teacher instructional quality is strongly connected to all other ways of measuring teacher quality (Bell et al., 2012). However, given the current policy environment in US schools, demonstrating a concurrent relationship with value-added scores is the accepted way to begin establishing the validity of any measure of teacher quality.

I also search for evidence that the validity of inferences might vary across different facets of measurement. In the analyses, I look at two separate measures of VA scores, a VA score constructed for the same year as observation scores were conducted (Current VA) and a

VA score constructed for the year before VA scores were conducted (Alt Year VA). Each measure presents some challenges for the validity of analyses.

V.4.1. Correlations with VA Scores across models In this section, I use the partial correlation between estimates of teacher quality from observation scores (i.e. $\hat{\upsilon}_t$) and value-added scores to provide evidence for the concurrent validity of observation scores, using this partial correlation as my "validity coefficient". Further, I test to see if this validity coefficient increases as I make adjustments for the effects of hidden facets. There are two reasons to think that the validity of observation scores will increase after adjusting for hidden facets. First, there was some evidence of instrument bias in the data presented to this point. If adjusting for hidden facets corrects for this instrument bias, then the validity of teacher estimates after making adjustments should increase. Second (and this reason applies mostly to the SO facets), the estimates of teacher quality could be biased by hidden facets as a result of co-construction effects (i.e. the hidden facet acts to increase or decrease observed teaching quality for all teachers), which could lead to biased estimates of teacher scores if those facets are ignored. In terms of the other effects that I have discussed in this chapter, those should affect estimates of the variance in teacher quality across teachers, but have no effect on the validity of the scores themselves.

Table 5.18 shows the partial correlation (i.e. “validity”) coefficients across the three instruments and the two VA measures. This table contains the results from many different regressions, displaying only the regression coefficient of interest. The columns indicate the model from which the estimate of teacher quality was drawn (all teacher quality estimates were standardized). The top two rows show the validity coefficient for CLASS, the next two for FFT, and the last two for PLATO. Within each set of rows, the top row (Alt Year VA) shows the results for the prior year's VA score estimate while the bottom row (Current VA) uses the current year VA estimate. The validity coefficients in Table 5.18 are all quite similar with minimal differences across VA score estimates or across different estimates of observed teacher quality, which is not surprising given the finding of high correlations across the teacher quality estimates from different models. With respect to differences across models that adjust for different facets, the validity coefficients using scores from the SO model were slightly lower than correlations for Mean, Base, SD, and CI models, but only for CLASS and FFT. However, the standard errors of the estimated parameters are larger than the differences in the validity coefficients across the different models. Thus, I cannot conclude that there are differences in the validity coefficients across the different models. This is likely due to the high correlations between teacher score estimates across observational models, which precludes the possibility of the validity coefficients from estimates differing very much (i.e. two variables that are correlated with each other at 1 will always have the same correlation with any third variable. When the two variables are correlated very close to 1, their respective correlations with any third variable must be almost the same).

Table 5.18: Partial Correlations between VA Scores and Teacher Quality Estimates across Models

Outcome	Mean ($E(X_{ir(o:d:s:t)})$)	Base ($\beta_{v_t^{base}}$)	SD ($\beta_{v_t^{SD}}$)	CI ($\beta_{v_t^{CI}}$)	SO ($\beta_{v_t^{so}}$)
CLASS					
Alt Year VA	0.21 (0.07)**	0.20 (0.08)**	0.21 (0.07)**	0.22 (0.07)**	0.17 (0.06)**
Current VA	0.19 (0.08)*	0.19 (0.08)*	0.17 (0.08)*	0.18 (0.08)*	0.14 (0.07)*
FFT					
Alt Year VA	0.14 (0.08)	0.16 (0.08)	0.17 (0.08)*	0.19 (0.08)*	0.14 (0.07)*
Current VA	0.21 (0.08)*	0.19 (0.08)*	0.19 (0.08)*	0.21 (0.08)**	0.16 (0.07)*
PLATO					
Alt Year VA	0.20 (0.07)**	0.18 (0.07)*	0.20 (0.07)**	0.24 (0.07)***	0.21 (0.06)***
Current VA	0.25 (0.07)***	0.23 (0.07)**	0.21 (0.07)**	0.26 (0.07)***	0.24 (0.06)***

Note. Table shows the estimated partial correlations between the indicated value-added score and estimated teacher quality from observation scores after controlling for student prior achievement and the demographic composite. Alt Year VA is the Prior Year Value-Added score estimate. Current VA is the Current Year Value-Added score estimate. Mean is the Teacher Quality score averaged from observed teaching quality. Base is the teacher score estimate from the Base model. SD is the teacher score estimate from the System Design model. CI is the teacher score estimate from the Curriculum and Instruction model. SO is the teacher score estimate from the School Organization model. Asterisks denote the significance of the relationship between the classroom observation score estimate and the VA score. * p<0.05; ** p<0.01; *** p<0.001.

V.4.2. Differential Validity across Facets In this last section, I look to see if the validity of inferences about teacher quality based on observation scores varies across the facets over which teachers were observed. As I argued before, there is no a prior reason to think that the relationship between observed teaching quality (i.e. $X_{ir(o:d:s:t)}$) and student learning is constant across different facets. Observation instruments may be better at measuring teacher quality for specific types of instruction or for specific types of students, for example. The results presented in this section follow on those in the last section. In this analysis, I introduce an interaction term into the regressions from the last section that used the Base model estimate of teacher quality. The interaction term is then examined to see whether the validity coefficient varies across levels of the hidden facet. In this sense, then, the interaction term is the main parameter of interest in the analysis. Note that, in these analyses, I aggregated all day-level facets to the teacher-level. As a result, in the analyses that follow, I am looking at whether teachers who were observed teaching more writing lessons have teacher quality scores that are more highly correlated to VA scores (and are hence more valid measures of teacher quality) than teacher quality scores from teachers who were observed teaching fewer writing lessons. This is not the ideal approach to examining concurrent validity since it does not address the potential impact that might arise from which teachers were observed teaching writing (i.e. teachers who are observed teaching writing more often may be fundamentally different in some way than those observed teaching writing less often). This would lead to the impression that scores on writing lessons are more valid than those on non-writing lessons, but the effect is driven by who was observed teaching writing, rather than writing itself. That is to say, this analysis is particularly exploratory and results should be verified before they are taken too seriously.

Table 5.19 shows the results for the analysis testing for differential validity across the grade facet (an SO variable). Each column of Table 5.19 shows the results of a different regression, with each row showing a regression coefficient. The last two rows, which show the interaction of grade and observation scores, are the focal parameters. A significant effect of these parameters suggests differential validity across the grade facet. The first three columns look at results for the current year VA scores while the next three look at results for the alternative year VA scores. There is no evidence that relationship between observation scores and VA scores varies across grade, as is shown in the bottom two rows.

Table 5.19: Regression Results Predicting Value-Added Scores with Observation Scores across Grade-Levels

	Current VA			Alt Year VA		
Parameter	CLASS	FFT	PLATO	CLASS	FFT	PLATO
Intercept	-0.13 (0.12)	-0.12 (0.11)	-0.13 (0.11)	0.08 (0.11)	0.08 (0.12)	0.08 (0.11)
Demo Composite	0.09 (0.10)	0.08 (0.10)	0.10 (0.10)	0.01 (0.10)	-0.02 (0.10)	0.02 (0.10)
Prior Ach	0.11 (0.11)	0.07 (0.11)	0.11 (0.10)	0.15 (0.11)	0.15 (0.11)	0.19 (0.11)
Grade 7	0.11 (0.17)	0.09 (0.17)	0.10 (0.17)	-0.14 (0.17)	-0.18 (0.17)	-0.17 (0.16)
Grade 8	0.20 (0.16)	0.19 (0.16)	0.18 (0.16)	-0.11 (0.15)	-0.10 (0.16)	-0.12 (0.15)
Obs. Score	0.12 (0.12)	0.17 (0.12)	0.28 (0.13)*	0.10 (0.11)	0.05 (0.12)	0.09 (0.13)
Obs Score by Grade 7	-0.01 (0.18)	-0.10 (0.17)	-0.19 (0.18)	0.10 (0.16)	0.04 (0.17)	0.10 (0.18)
Obs Score by Grade 8	0.23 (0.16)	0.18 (0.16)	0.05 (0.16)	0.19 (0.16)	0.17 (0.16)	0.15 (0.16)

Note. Cells show the regression parameters with SE. Each column is a separate regression. Demo Composite=Demographic Composite; Prior Ach= Prior Achievement; Obs Score= Observation Score-Estimated Teacher Quality from Base Model. * p<0.05; ** p<0.01; *** p<0.001.

Table 5.20: Regression Results Predicting Value-Added Scores with Observation Scores across Student Characteristics

	Current VA		Alt Year VA	
Parameter	Pr Ach	Demo	Pr Ach	Demo
CLASS				
Intercept	0.00 (0.07)	0.03 (0.07)	0.05 (0.07)	0.02 (0.07)
Demo Composite	0.09 (0.10)	0.11 (0.10)	0.04 (0.10)	0.04 (0.11)
Prior Ach	0.11 (0.11)	0.12 (0.11)	0.18 (0.11)	0.17 (0.11)
Obs. Score	0.20 (0.08)*	0.21 (0.08)**	0.23 (0.08)**	0.22 (0.08)**
Obs Score by Facet	-0.04 (0.07)	0.10 (0.08)	-0.11 (0.07)	0.06 (0.08)
FFT				
Intercept	0.02 (0.08)	0.05 (0.07)	0.07 (0.07)	0.03 (0.07)
Demo Composite	0.07 (0.10)	0.08 (0.10)	-0.00 (0.10)	-0.01 (0.10)
Prior Ach	0.07 (0.11)	0.08 (0.11)	0.16 (0.11)	0.15 (0.11)
Obs. Score	0.21 (0.08)**	0.22 (0.08)**	0.15 (0.08)	0.15 (0.08)
Obs Score by Facet	-0.05 (0.07)	0.12 (0.07)	-0.13 (0.07)*	0.07 (0.07)
PLATO				
Intercept	-0.03 (0.07)	0.01 (0.07)	0.00 (0.07)	-0.03 (0.07)
Demo Composite	0.09 (0.10)	0.10 (0.10)	0.04 (0.10)	0.02 (0.10)
Prior Ach	0.10 (0.10)	0.10 (0.10)	0.19 (0.11)	0.18 (0.10)
Obs. Score	0.25 (0.07)***	0.27 (0.07)***	0.20 (0.07)**	0.18 (0.07)*
Obs Score by Facet	0.02 (0.08)	0.08 (0.07)	-0.04 (0.07)	-0.05 (0.07)

Note. Cells show the regression parameters with SE. Each column and block is a separate regression. Demo Composite=Demographic Composite; Prior Ach= Prior Achievement; Obs Score= Observation Score-Estimated Teacher Quality from Base Model; In the second and fourth columns, Facet is 'Pr Ach'=Prior Achievement; In the third and fifth columns, Facet is Demo=Demographic Composite. * p<0.05; ** p<0.01; *** p<0.001.

Table 5.20 shows the results for the student demographic composition (another SO variable), presenting equivalent information to that of Table 5.19. Each column of Table 5.20 shows the results of a separate regression for each of the three observation instruments. The left two columns show results for the current year VA scores and the right two for the alternate year VA scores. In only one of the twelve equations in Table 5.20 is there evidence that the relationship between estimates of teacher quality and VA scores varies across levels of this facet. For the alternate year VA score only, the relationship between observation scores and VA scores is weaker for teachers with more disadvantaged students on FFT. However, this interaction would not be significant after correcting for multiple comparisons.

Table 5.21 shows the same analysis for the content domain facet (a CI variable). Again, each column of Table 5.21 shows the results of a separate regression for each content domain facet on each observation instrument. Only one interaction (out of 24) is significant here. Teachers who were observed teaching more reading lessons had a higher validity coefficient for the alternate year VA score on FFT compared to those observed teaching fewer reading lessons. Again, the effect would not be significant after adjusting for multiple comparisons. Table 5.22 shows the same results for interaction structure. No effects were significant here.

Overall, the results presented in this section show that estimates of teacher quality across all three instruments are related to VA scores, but there is no evidence for differential validity across different levels of the hidden facets analyzed here. This means that there is no evidence that teacher quality is better measured when observing specific forms of instruction compared to other types of instruction. However, to really examine the differential validity across facets, it would be ideal to get stable teacher quality estimate for each level of the facet. This would require sampling multiple days of instruction for each teacher on each level of the facet. With separate teacher quality estimates for each level of the hidden facet, one could explore the within-teacher differences in the relationship of the two teacher quality estimates and VA scores. This should both eliminate the threat that non-teacher sources are creating the observed relationship and, since all teachers have reliable estimates of teacher quality for each facet, should provide more power to detect differences in the correlations with VA scores.

Table 5.21: Regression Results Predicting Value-Added Scores with Observation Scores across Content Domains

	Current VA				Alt Year VA			
Parameter	Read	Lit	Writ	Grammar	Read	Lit	Writ	Grammar
CLASS								
Intercept	0.00	0.08	0.06	-0.17	0.03	0.05	0.07	-0.16
	(0.08)	(0.09)	(0.10)	(0.09)	(0.07)	(0.09)	(0.10)	(0.09)
Demo Composite	0.09	0.08	0.10	0.10	0.03	0.01	0.01	0.04
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Prior Ach	0.10	0.13	0.11	0.12	0.13	0.16	0.14	0.17
	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.10)
Facet	-0.05	-0.08	-0.07	0.14	-0.11	-0.08	-0.07	0.14
	(0.12)	(0.08)	(0.08)	(0.07)*	(0.11)	(0.08)	(0.07)	(0.07)*
Obs. Score	0.20	0.29	0.13	0.30	0.16	0.16	0.23	0.32
	(0.09)*	(0.11)**	(0.11)	(0.10)**	(0.08)	(0.10)	(0.11)*	(0.10)***
Obs Score by Facet	-0.01	-0.07	0.06	-0.09	0.18	0.05	-0.03	-0.10
	(0.12)	(0.07)	(0.08)	(0.07)	(0.11)	(0.06)	(0.08)	(0.07)
FFT								
Intercept	0.02	0.08	0.07	-0.16	0.02	0.03	0.08	-0.15
	(0.08)	(0.09)	(0.10)	(0.09)	(0.07)	(0.09)	(0.10)	(0.09)
Demo Composite	0.06	0.05	0.07	0.08	-0.02	-0.03	-0.02	-0.01
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Prior Ach	0.05	0.08	0.08	0.08	0.14	0.15	0.14	0.16
	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.12)	(0.11)
Facet	-0.06	-0.05	-0.07	0.15	-0.08	-0.06	-0.07	0.13
	(0.11)	(0.08)	(0.08)	(0.07)*	(0.11)	(0.08)	(0.08)	(0.07)
Obs. Score	0.20	0.32	0.14	0.27	0.06	0.09	0.12	0.25
	(0.09)*	(0.10)**	(0.12)	(0.10)**	(0.09)	(0.10)	(0.13)	(0.10)*
Obs Score by Facet	0.05	-0.11	0.05	-0.03	0.22	0.06	0.01	-0.11
	(0.11)	(0.07)	(0.08)	(0.07)	(0.11)*	(0.07)	(0.08)	(0.07)
PLATO								
Intercept	0.01	0.09	0.10	-0.21	0.02	0.05	0.09	-0.18
	(0.07)	(0.09)	(0.10)	(0.09)*	(0.07)	(0.09)	(0.10)	(0.09)*
Demo Composite	0.09	0.08	0.10	0.12	0.03	0.03	0.03	0.05
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Prior Ach	0.09	0.13	0.11	0.13	0.17	0.21	0.20	0.21
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.11)	(0.11)	(0.10)*
Facet	-0.10	-0.12	-0.11	0.17	-0.13	-0.07	-0.10	0.15
	(0.11)	(0.08)	(0.07)	(0.06)**	(0.11)	(0.08)	(0.07)	(0.07)*
Obs. Score	0.22	0.30	0.19	0.32	0.16	0.19	0.08	0.31
	(0.08)**	(0.10)**	(0.11)	(0.10)***	(0.08)*	(0.10)	(0.11)	(0.09)***
Obs Score by Facet	0.14	-0.02	0.07	-0.03	0.17	0.02	0.11	-0.09
	(0.12)	(0.07)	(0.08)	(0.07)	(0.12)	(0.07)	(0.08)	(0.07)

Note. Cells show the regression parameters with SE. Each column and block is a separate regression. Demo Composite=demographic composite; Prior Ach= prior achievement; Obs Score= observation score-estimated teacher quality from Base model; In the second and sixth columns, the facet is days observed teaching reading; In the third and seventh columns, the facet is days observed teaching literature; In the fourth and eighth columns, the facet is days observed teaching writing; In the fifth and ninth columns, the facet is days observed teaching grammar. * p<0.05; ** p<0.01; *** p<0.001.

Table 5.22: Regression Results Predicting Value-Added Scores with Observation Scores across Interaction Structures

	Current VA			Alt Year VA		
Parameter	Discussion	Ind Wk	Recitation	Discussion	Ind Wk	Recitation
CLASS						
Intercept	0.13 (0.16)	0.06 (0.08)	0.01 (0.24)	0.30 (0.15)*	0.02 (0.08)	-0.11 (0.24)
Demo Composite	0.08 (0.10)	0.09 (0.10)	0.08 (0.10)	0.01 (0.10)	0.03 (0.10)	0.02 (0.10)
Prior Ach	0.11 (0.11)	0.11 (0.11)	0.11 (0.11)	0.17 (0.10)	0.16 (0.11)	0.14 (0.11)
Facet	-0.07 (0.07)	-0.16 (0.11)	-0.01 (0.08)	-0.14 (0.07)*	-0.05 (0.11)	0.04 (0.08)
Obs. Score	0.24 (0.17)	0.25 (0.09)**	0.66 (0.29)*	0.44 (0.15)**	0.24 (0.09)**	-0.14 (0.28)
Obs Score by Facet	-0.01 (0.07)	-0.10 (0.11)	-0.16 (0.10)	-0.10 (0.06)	-0.06 (0.10)	0.12 (0.09)
FFT						
Intercept	0.15 (0.16)	0.07 (0.08)	0.08 (0.25)	0.27 (0.15)	0.02 (0.08)	-0.12 (0.24)
Demo Composite	0.05 (0.10)	0.07 (0.10)	0.06 (0.10)	-0.04 (0.10)	-0.02 (0.10)	-0.03 (0.10)
Prior Ach	0.06 (0.11)	0.06 (0.11)	0.06 (0.11)	0.16 (0.11)	0.14 (0.11)	0.13 (0.11)
Facet	-0.07 (0.07)	-0.18 (0.11)	-0.02 (0.08)	-0.13 (0.07)	-0.04 (0.11)	0.04 (0.08)
Obs. Score	0.23 (0.17)	0.23 (0.09)*	0.62 (0.28)*	0.26 (0.16)	0.15 (0.09)	-0.14 (0.27)
Obs Score by Facet	-0.00 (0.08)	-0.01 (0.11)	-0.14 (0.09)	-0.05 (0.07)	-0.01 (0.10)	0.10 (0.09)
PLATO						
Intercept	0.23 (0.16)	0.06 (0.08)	-0.01 (0.24)	0.36 (0.15)*	0.01 (0.08)	-0.10 (0.24)
Demo Composite	0.09 (0.10)	0.10 (0.10)	0.09 (0.10)	0.02 (0.10)	0.03 (0.10)	0.03 (0.10)
Prior Ach	0.12 (0.10)	0.11 (0.10)	0.10 (0.10)	0.20 (0.10)*	0.19 (0.11)	0.18 (0.11)
Facet	-0.12 (0.07)	-0.19 (0.11)	-0.00 (0.08)	-0.17 (0.07)*	-0.05 (0.11)	0.03 (0.08)
Obs. Score	0.36 (0.15)*	0.27 (0.08)**	0.47 (0.23)*	0.35 (0.14)*	0.20 (0.08)*	0.01 (0.22)
Obs Score by Facet	-0.04 (0.07)	-0.01 (0.12)	-0.08 (0.08)	-0.05 (0.06)	0.01 (0.11)	0.07 (0.07)

Note. Cells show the regression parameters with SE. Each column and block is a separate regression. Demo Composite=Demographic Composite; Prior Ach= Prior Achievement; Obs Score= Observation Score-Estimated Teacher Quality from Base Model; In the second and fifth columns, Facet is Days with a sustained focus on discussions; In the third and sixth columns, Facet is Days with a sustained focus on independent work; In the fourth and seventh columns, Facet is Days with a sustained focus on recitation. * p<0.05; ** p<0.01; *** p<0.001.

V.5. Chapter Summary

In this chapter, I reviewed a number of results about the effect of facets (and especially "hidden" facets") of measurement on observed teaching quality and the resulting implications for bias, reliability and validity. To address my first research question, I started by exploring the facets of measurement typically built into the design of most observation systems and explicitly included in most GTheory models. These facets include teachers, occasions, days, raters, and items. Teachers contributed less to observed teaching quality on PLATO than the other two instruments and teachers contributed the most to observed teaching quality on FFT. Additionally, the rater facets contributed a large portion of the variance in observed teaching quality, especially the rater-by-day and rater-by-item facets. For CLASS in particular, the variance of the rater-by-item facet was large. Item facets also made a large contribution to observed teaching quality, especially the item-by-day facet, though items contributed relatively little variance to observed teaching quality at the teacher level (i.e. item-by-teacher facets were small). These facet variances were estimated with a surprisingly high degree of precision (with confidence intervals spanning a few percentage points), but the error was large relative to the size of the variance components, which contributed to the high error in reliability estimates we saw in the latter part of this chapter.

In order to address my second research question, I presented the results of three models that build off of this Base model to add controls for the impacts of the hidden facets under study on observed teaching quality. These models showed that SD, CI, and SO facets all were associated with observed teaching quality. These effects were often consistent across instruments, but for live scoring, reading, grammar, discussion, the student demographic composite, and student's average prior achievement, there was a significant difference between the effects of the hidden facets on observed teaching quality across instruments, which suggests instrument bias. This bias appeared to be the result of the specific aspects of teaching quality captured by each instrument, with each instrument capturing a unique component of teaching quality.

However, despite these effects on observed teaching quality, I showed that the impact of the hidden facets on estimates of teacher quality were often quite small, except for the effects of the SO facets. This is likely due to the random sampling of days, which averaged out the impact of day-level facets (like the SD and CI facets) on estimates of teacher quality. Even the effects of the SO facets on estimates of teacher quality were modest. I showed that the impact of adjusting for the hidden facets on the reliability of teacher quality scores was more meaningful. While the decrease in reliability was modest after adjusting for the SD and CI facets, it was large enough to suggest adding an additional day of observation for each teacher, which entails a large cost. The decrease in reliability after adjusting for the SO facets was much more substantial with the reliability of teacher quality scores from the SO model remaining below 0.55 even when five days of instruction are observed and half of the observed days are double scored. Overall, then, this chapter: (a) provides convincing evidence for the presence of effects of hidden facets; (b) shows that adjusting estimates of teacher quality for hidden facets has relatively little effect on teacher score estimates, at least in UTQ data, and (c) suggests the design of the observational system should depend on whether adjustments for hidden facets will be made to estimates of teacher quality.

I then presented results related to my third research question, the differential validity of estimates of teacher quality. These results showed that there was no difference in the correlation of the estimates of teacher quality with VA scores across the different models (i.e. the Base model, SD model, CI model, and SO models). Further, the correlation with teacher quality estimates did not vary for teachers observed across different facets. This showed that there was no evidence for the differential validity of inferences of teacher quality.

Chapter VI. Discussion

VI.1. The Problem

This thesis explored the implications of treating teaching as a situated phenomenon for the measurement of teacher quality with classroom observation instruments. Traditional approaches to measuring teaching quality using classroom observation instruments have recognized that observed teaching quality scores will vary across days, class sections taught by a teacher, particular items on an observation instrument, and the raters using the observation instrument. However, researchers also typically assume that each level of these facets provides an equivalent view of teaching quality. As such, the typical approach to measurement implicitly assumes that any two days of instruction are an equal representation of a teacher's ability to mount high quality instruction, that any two raters provide equally valid scores, and so on.

The problem addressed in this thesis is what happens when variation in teaching quality occurs in systematic ways across days, raters, and items. For example, suppose that certain properties of days—for example, the teacher's use of lecture or class discussion—has a systematic effect on teaching quality. When this occurs, days featuring lectures are not equivalent to days featuring discussions. Further, a teacher's ability may not be measured as accurately if observed only during lectures as compared to both lectures and class discussions. When specific days (or levels within any other facet) are systematically related to teaching quality, measurement models may provide incorrect parameter estimates. Characteristics of measurement that are systematically related to teaching quality, but not explicitly included in measurement models, are called "hidden" facets of measurement. The question explored in this thesis was whether we can identify some of these hidden facets, and if so, how explicitly incorporating these hidden facets into our measurement analysis affects the bias, reliability, and validity of our inferences about teacher quality.

The presence of hidden facets raises several issues in the measurement of teaching quality. The first involves the problem of generalizing and/or extrapolating scores in the face of these hidden facets. I argued in this thesis that when a facet varies within-teachers and teachers are observed across representative levels on this facet, it is a straightforward matter to understand how observed teaching quality generalizes across levels of the facet because there is data on this issue. However, when a facet varies mostly between-teachers and teachers are only observed on a small part of the domain of this facet, it is much more difficult to generalize because we do ***not*** have direct data on how observed teaching quality varies across that facet (for a given teacher). In this case, generalization involves a certain amount of extrapolation.

Importantly, the distinction between within- and between-teacher facets may be sample and observation protocol dependent. For example, the UTQ study data used in this thesis came from a study design that sampled teachers separately in math and English, making subject a between-teacher facet; but other studies (often in elementary schools), might sample math and English lessons from the same teacher, making subject a within-teacher facet. When a facet is between-teachers, some assumptions must be made about how teachers observed across the facet differ in order to generalize across the domain of the facet. This is a process of extrapolation because generalization is accompanied by a set of assumptions (which may or may not be true) about the nature of teacher-to-teacher differences across levels of the facet and these assumptions must be true for generalization to be accurate. The problem of this thesis, then, comes down to understanding the facets of measurement, which determine the boundaries across which generalization occurs, and to determining how observed teaching quality varies across these facets, including the assumptions necessary to generalize across facets where extrapolation is necessary.

As I discussed throughout this thesis, the problem of generalization first involves identifying potential "hidden" facets of measurement that affect observed teaching quality and over which generalization is desired. The first category of hidden facets that I studied was System Design (SD) variables. These are characteristics of classroom observation systems that arise as part of the selection of specific days to observe, raters to conduct observations, and procedures to score observation data. Among the variables considered in this study, for example, were the time of year and day of the week when data were video recorded, the date in the study period when video scoring occurred, and whether or not scores were recorded live or from video data. As discussed in this thesis, SD facets like the ones studied here are usually within-teacher facets because teachers are usually observed across a wide range of levels on these facets (e.g. teachers are observed during set observation windows spread across the school year, and their videos are scored from video at many time points across the study period). Because of this, data are sufficient to generalize observed teaching quality across these facets so that a simple averaging of a teacher's scores across all observation occasions will typically result in a reasonably unbiased score, assuming the observation protocol was well-designed and implemented. However, this averaged score will contain not only a true score component of variation but also the variance in observed scores due to the omitted hidden facets. This was demonstrated in this thesis by building a GTheory statistical model that statistically adjusted for the SD facets. This model eliminated variation in teacher scores due to SD variables (like when teachers were observed or whether a teacher's videos were scored live or by video), and as we saw, this statistical adjustment had implications for the reliability of scores (by reducing the ratio of "true" score variance to error variance). It did not, however, have much effect on point estimates of teacher quality since the random sampling of days and the random assignment of raters across days in UTQ already balanced the effects of these facets.

The second category of hidden facets that I studied involved dimensions of Curriculum and Instruction. In this thesis, the Curriculum and Instruction (CI) facets studied included the structure of instructional interactions occurring during observed lessons as well as the ELA content domains that were taught. Because most observation protocols (including UTQ) sample days of instruction more or less randomly, teachers tend to be observed across a range of these CI facets (e.g. they are observed teaching writing some days and reading other days). Thus, CI facets (like the SD facets) are within-teacher facets. However, there may be between-teacher components to CI facets if, for example, a writing curricula in some schools leads writing instruction to be fundamentally different in some schools than others (which makes the writing instruction facet between-teachers since a teacher is observed only in one style of writing instruction)[52]. Alternatively when different teachers teach the CI facets with different frequencies, between-teacher effects of the CI facets may exist.

Models built to statistically adjust for the CI facets (such as the CI model used in this thesis) estimate teacher quality within each of the CI facets, which eliminates differences in teacher quality stemming from how frequently teachers teach a given topic. Therefore, the "adjusted" teacher quality estimate from the CI model captures a teacher's ability to teach reading and writing ***not*** the frequency of teaching it. Importantly, there is a bias-variance trade-off in this adjustment decision. The negative impact of introducing a bias by ignoring aspects of teacher quality linked to the frequency with which teachers engage in specific types of instruction can be outweighed by the benefit of reducing sampling variation

[52] In the UTQ data, I found between-teacher effects of a CI facet (namely content domain taught), but, as I argued before, this does not necessarily make the CI facet a between-teacher facet because teachers were observed across a representative range of the facet's domain due to the random sampling of days. The distinction of within-teacher and between-teacher facets stems more from whether data exists for generalization, not from whether they affect between-teacher differences in observed teaching quality.

stemming from how frequently teachers are observed engaging in specific types of instruction. I have argued that too few days of instruction are observed to estimate how often a teacher teaches writing (or any other CI facet) so the reduction in sampling error will likely outweigh the introduction of bias. However, this trade-off is likely sample dependent (i.e. dependent on the relative size of within-teacher and between-teacher effects of CI facets) and hard to evaluate. Importantly, this tradeoff only occurs under a limited set of conditions—when it is difficult to accurately estimate the frequency with which a teacher teaches at particular levels of the facet *and* when the facet has (between-teacher) effects on teaching quality. Thus, the precision-bias trade-off likely exists only for some CI facets and for some ways of defining and understanding teacher quality. When the trade-off does not exist, adjusting for the facet will increase precision without negative effects (i.e. without affecting bias).

The third category of hidden facets that I studied were called School Organization (SO) facets. In this thesis, SO facets included features related to the design of school systems like student composition, grade taught, and subject taught. As discussed earlier in this thesis, SO facets are usually between-teacher facets, so generalizations from observations on a given teacher to other settings can sometimes involve extrapolation. For example, if we observe a teacher teaching students with only a limited range of background characteristics, we must extrapolate in order to compare that teacher's measured quality to the measured quality of teachers who teach students with an entirely different range of backgrounds. In models that do not directly adjust for the SO facet effects, differences in measured teaching quality due to context get attributed to teachers. This would be a good procedure if, in fact, differences in averaged teaching quality across contexts was due to teacher sorting (i.e. differences in the ability of teachers who are employed at different levels of the facets). On the other hand, we might assume that facet effects arise not from teacher sorting but from co-construction (i.e.

the facet itself causes differences in observed teaching quality due, for example, to some students being easier to teach than others). In this case, scores are only comparable across contexts after adjusting for the effects of the SO facets. Importantly, in both cases, extrapolation is necessary if one wants to generalize beyond the specific setting where a teacher was observed because teachers were not observed across a representative range of the domain of SO facets. The point, once again, is that assumptions drive the way a teacher quality score is estimated, as well as the extent to which one can generalize this score to settings which have not been directly observed *and* these assumptions should be made clear along with the goals of generalizing. Moreover, it will always be the case that teacher quality estimates based on adjusted and unadjusted models will produce somewhat different point estimates (to the extent that SO facets have effects on observed teaching quality) and vary in the precision of their estimates.

The division of contextual features of instruction into within-teacher and between-teacher hidden facets and the categorization of three types of hidden facets provided a framework to explore how the situated nature of teaching affects the measurement of teacher quality with observation instruments. This is one of the major contributions of this thesis because it allows for an exploration of how contextual factors of instruction (i.e. hidden facets) impact the measurement process, focusing on what we can conclude about teacher quality and the limitations of different estimates of this construct.

VI.2. Review of Findings

Having discussed the problem of generalization in the face of hidden facets, I turn now to a review of the findings. I begin with a discussion of how planned facets of measurement (such as occasions, days, raters, and items) affect the measurement of teacher quality.

VI.2.1. Planned Facets of Measurement

The analyses I presented in this thesis began with a presentation of one of the most complete GTheory analyses conducted to date on the effects of planned measurement facets on observed teaching quality scores for the observation instruments under study (note that the analysis also broke new ground by presenting confidence intervals to bound estimated variances in this analysis). The GTheory model that I estimated (called the Base model in previous chapters) produced a number of interesting findings. The first was that the amount of variance in observed scores due to the teacher facet (i.e. the true score, υ_t) differed across the three instruments under study. The teacher facet contributed the most variance to observed scores on FFT and the least to observed scores on PLATO. In these initial analyses, then, FFT was found to provide the most reliable estimate of teacher quality and PLATO was found to provide the least reliable estimate. Note, however, that other studies—including the MET study (Kane et al., 2012)—of these same instruments have not found large differences in reliability across instruments. Note also that the differences I found in estimated score reliability across the three instruments under study did not seem to affect the concurrent validity of these measures (i.e., the correlation of the estimated teacher quality scores to VA scores was roughly similar for all three instruments).

An interesting contribution of this study—and one not found in other GTheory analyses of these observation instruments—was my calculation of confidence intervals for each of the variance estimates in my model, which allowed me also to calculate confidence intervals for the variance of planned measurement facets. The 95% confidence intervals for the variance estimates were large relative to the estimates themselves, but bounded the percentage of variance explained by a facet to +/- ~3 percentage points. This large error relative to the variance components themselves contributed to wider than desired confidence intervals on estimates of score reliability, making it hard to determine whether reliability changed significantly as planned facets of measurement (such as number of days or number of raters) changed. In any case, it is important to consider what the confidence intervals I estimated do and do not represent. Importantly, researchers who want to make use of the findings presented here to plan their own studies will have to extrapolate these findings to their own setting. The confidence intervals do *not* show the uncertainty likely to arise in this extrapolation process, but show uncertainty in running a similar study in the same context. At present, no research has explored the limits or boundaries of this extrapolation process, though the similarity in the relative sizes of facet variances across studies of different populations (when common statistical models are used) suggests that generalizing across fairly similar populations might be warranted.

A further caution about the confidence intervals constructed here is warranted. The bootstrapping method I used to construct confidence intervals assumes that model estimates of variance components do, in fact, reflect population parameters. This is the basis from which re-sampling is used to calculate uncertainty in parameter estimates. I have expressed concerns, which I review again below, that the structure of the UTQ data, assignment of raters, and relatively low rate of double scoring puts limitations on the estimation of some variance components. If that is the case, the bootstrapped confidence intervals may not be correct. Simulation studies (or the computationally prohibitive double bootstrap) could be used in the future to test how the complex structure of the data might affect estimated variance components.

The most unique feature of the Base model that I estimated was the inclusion of items as an explicit (i.e. planned) facet of measurement. The item models presented in this thesis provided a great deal of useful information about the functioning of observation instruments as measurement tools. Across all three instruments, item fixed effects were always large although noticeably more so for CLASS and PLATO than for FFT. These large item effects suggest the need to better model how items function individually (Shavelson et al., 1986).

That is, my models assumed each item varied across days, teachers, and raters to the same extent, which may not be true and should be explicitly tested. The item-level models in Appendix D do this, but an in depth exploration of these models is beyond the scope of this thesis. The item-by-rater and item-by-day interactions show further the importance of considering items in a standard GTheory model of classroom observation instruments. Items, apparently, do not have a consistent "difficulty" across raters or days. While future studies might try to explain why this is the case, in this thesis, I simply treated these item interactions as error in the measurement process.

The Base GTheory model used in this thesis also went beyond the typical analysis of rater error found in classroom observation research. In many studies, the only rater effect estimated is rater main effects (e.g. Cor, 2011; McCaffrey, et al., 2014). But I also estimated various rater interaction effects. These interaction effects, in turn, showed the importance of rater error other than simple leniency (as estimated by the rater main effect). In particular, the Base model I estimated showed substantial rater-by-item interactions, suggesting that raters are not consistently lenient (or severe) across all items, and rater-by-day interactions, suggesting inconsistencies in rater leniency across days. In fact, in the Base models estimated here, the rater-by-day, rater-by-item, and residual facets were always a large source of error variance, but the relative magnitude of these error components varied across instruments. In fact, this variation across instruments has important implications for efforts to reduce rater error. For example, the high rater-by-item error on CLASS, which accounted for almost one-fifth of the total score variance, shows that raters struggled with understanding the level of teaching quality that corresponds to a specific score value on a given item. On FFT, however, the rater-by-day error was the largest by far, showing that (the same) raters struggled the most with understanding the level of teaching quality being exhibited on specific days of instruction. While no research exists to connect these error types with specific training remedies to reduce rater error, it is reasonable to think that different approaches would be required to address each of these two very different types of rater error. This deeper understanding of rater error, which hopefully will better guide researchers to solutions, is one of the benefits of including five sources of rater error in the GTheory model. However, more research is needed to connect the various types of rater error with approaches to reducing these errors. One final, cautionary note on rater error is essential. At least in UTQ, up to half the rater error may be undetectable because no "true" observed teaching quality is available to judge rater error. With the data at hand, we can only examine rater disagreement (Myford & Wolfe (2009); White, In Prep). So, when two raters are both wrong in the same way, we must incorrectly conclude they are correct.

VI.2.2. The Relationship of Data Structure, Rater Errors, and Score Reliability

In estimating the Base (and other) GTheory statistical models, I raised concerns about how the complexity of the UTQ data structure might have affected the results presented here. For example, the rater-by-teacher and rater-by-day facets were not well-separated (their variance estimates are correlated at ~ -0.8 across bootstrapped samples), and this could be due to the complex UTQ data structure, which includes only a partial crossing of raters (i.e. all occasions of instruction are not scored by all raters). The need for a partial crossing of raters is obvious (it is too costly to have all raters score all occasions), but that need forces those conducting a measurement study to make study design decisions that probably affect the estimation of variance due to rater facets. To begin, any partial crossing of raters involves choosing a level of nesting at which to assign raters. That is, raters must either be assigned to teachers, such that they score all days for a given set of teachers, be assigned to days, such that they score all occasions for a given day (and a limited number of days per teacher), or assigned to occasions, such that they score only one occasion for a given day. Second, a decision must be made about whether to score occasions sequentially or independently. Generally speaking, the lower the level of nesting to which raters are assigned (i.e. occasions rather than days), the more raters will contribute to a teacher's observed score because fewer raters will contribute to each teacher’s score. Assigning raters to occasions will lead to the most reliable teacher scores because the rater, rater-by-item, and rater-by-day errors will be spread across the most raters. However, this approach would also mean that fewer raters were scoring the same occasions on a given day or the same days within teachers, reducing the power to detect systematic rater biases (e.g. raters who are biased against lectures or minority teachers). There is thus a trade-off here: increasing the reliability of estimates of teacher quality or increasing ability to explore rater errors. The Base GTheory model allows an exploration of this trade-off. The rater-by-teacher error facet was near zero in the analyses I conducted, so adding additional raters to score a given teacher does not contribute to score reliability. The rater-by-day error facet, however, was large, so adding more raters to score each day is an important step to increasing reliability[53]. It would seem, then, that raters should be assigned to the occasion level to maximize the reliability of scores since this maximizes the number of raters scoring each day. Alternatively, if one wishes to explore rater biases, assigning raters to the teacher level is preferable, but this will increase error in teacher score estimates and should probably only be done if at least two raters score each teacher. Additionally, there seems to be no statistical reason to ever assign raters to the day level since this hampers exploration of rater bias and provides less reliable scores than assigning raters to occasions. Interestingly, most studies assign raters to days. This may reflect the importance of non-statistical reasons. For example, one may be concerned that a rater cannot score the second occasion without viewing the first occasion for context or scoring may be done live, in which case assigning raters to occasions is impractical.

[53] The rater and rater-by-item facets would seem to play a role here, but given sufficient double scoring, a regression model with rater-by-item fixed effects (i.e. include a dummy variable for every rater, every item, and every combination of rater and item in the regression and use the residual from this model as observed scores) adjusts for these effects, preventing them from influencing score reliability for any fixed set of raters. If we want to generalize effects beyond the observed raters, this approach is not possible and having more raters score each teacher will reduce the effect of these facets.

VI.2.3. How important are hidden facets? The Base model just discussed was not the focus of this dissertation. Instead, the main goal of the thesis was to explore the role that hidden facets play in the measurement of teaching quality. Three categories of hidden facets were studied in my thesis—SD facets, CI facets, and SO facets. Notably, each category of hidden facets contained variables that had a statistically significant effect on observed teaching quality, but the addition of the SD and CI facet effects into the model led to almost no differences in estimates of teacher quality. This was because SD and CI facets are within-teacher facets and SO facets are between-teacher facets. Differences in teacher quality estimates across the Base and CI model (which adjusted for all within-teacher hidden facets) was very close to 1 because these day level, within-teacher, hidden facet effects were averaged across four days, sampling of days was random (or at least ignorable), and the UTQ study stratifies teacher scores across SD facets. This finding might be sample dependent as the UTQ data showed much less day variance than previous studies (such as MET, see Kane et al., 2012).

Going forward, it is worth exploring whether the within-teacher hidden facets have larger effects in practical evaluation applications, where sampling is less controlled and typically under teacher or principal control. In fact, because SD and CI facets operate within-teachers, teachers can "game the system" by making selected decisions about when they will be observed (e.g., when in the day or year, or teaching a reading versus a writing lesson). The large size of the effects of the hidden facets on observed teaching quality implies that these decisions will have major implications for teacher scores, allowing teachers to significantly move up the distribution of teacher quality by cleverly "gaming" the system. For example, a teacher that is able to be observed *only* at the beginning of the school year on writing lessons that feature discussions in their 6th grade class will score in the 98th percentile of teacher quality on CLASS while this same teacher, had they been observed *only* towards the end of the school year, in their 7th grade classroom, and on reading lessons that did not feature discussions, would have been scored at the 50th percentile. While this example is admittedly an extreme case, it shows the potential control a well-informed teacher can have on their scores by controlling when they are observed. The best solution to this challenge is to keep sampling as close to random as possible.

However, near random sampling is not possible, nor necessarily always desirable in practice. For example, a teacher may wish to be observed only in writing because they need formative feedback on their writing instruction (while other teachers are observed only in grammar for similar reasons). Allowing this may be beneficial to the formative feedback goals of observation systems, but still has the same result just discussed: the within-teacher effects found in the UTQ data take on between-teacher components and lead to larger differences in teacher quality estimates across models that make different adjustments for hidden facets. In this way, there can be a tension between formative and summative uses of observational systems. Thus, there is a need to take care in generalizing the results of this study to practical applications. More research is needed to examine how the sampling of lessons in practice might affect scores, especially when teachers face both high-stakes consequences and have some control over what days are observed (Brophy, 2006).

While the within-teacher (SD and CI) hidden facets had minor impacts on teacher score estimates overall in the UTQ data, they did have important (though modest) effects on the reliability of teacher score estimates. The estimated variance of the teacher score fell by 21-38% between the Base model and the CI model across the three instruments. This implies that 21-38% of the variance in teacher scores from the Base model is the result of sampling error due to sampling across the SD and CI facets. This is a lot of "error" in the teacher score estimates from the Base model, though, as I have discussed, terming this error depends which aspects of teacher quality we want to include in our definition of teacher quality. Nonetheless, the finding about large error variance due to hidden facets suggests that estimates of teacher quality are not as reliable as the Base model would suggest. The decrease in the estimated reliability of the teacher score dropped enough for FFT and PLATO to suggest that an additional day of observation is necessary to maintain the same level of reliability indicated by the Base model, which has important implications for the cost of using classroom observation instruments. However, it is once again worth noting that the uncertainty in estimates of the reliability of teacher scores is much larger than differences across models, an important point given that no previous studies have investigated the precision with which these reliabilities are estimated. At least for UTQ data, then, we can conclude that the effects of within-teacher hidden facets are large enough to be theoretically of interest and are helpful in understanding the reliability of teacher scores, but that inclusion of these hidden facets in a GTheory model has little practical effect on teacher score estimates themselves. This changes when sampling is not well-controlled as non-random sampling can lead to the hidden facets having much larger effects on the teacher score estimates.

The effects of the between-teacher facets, namely the SO facets, are a different story. These facets are unique in that comparisons of teacher scores across between-teacher facets are only supported through extrapolation. As a result, understanding the source of these effects and determining how to address them is highly complex. Models that adjust for the SO facets resulted in estimates of teacher quality that, though still highly related to the Base model estimates (correlations were near 0.8), were noticeably different. While a correlation of .8 is still quite high, the implications could differ. Further, estimates of the reliability of scores fell roughly 0.15-0.20 points from the Base model to the SO model, a decrease that is easily large enough to imply additional raters or days of instruction are necessary to achieve a given score reliability. In fact, under common sampling plans, the reliability estimates from the SO model suggest very little ability to reliably distinguish between teachers, which is a primary goal of observation instruments in most research and practice settings. This finding, in fact, is the major benefit of incorporating this exploration of hidden facets within a broader measurement framework, which is rarely done.

This review of findings from the SO models raises the important question of whether or not it is appropriate to make adjustments for student characteristics and grade taught when estimating teacher quality from classroom observation data. It cannot be determined from UTQ data whether teacher sorting or co-construction were the source of the SO facet effects found in this study, which means I cannot conclude definitively if an adjustment for SO effects is appropriate.

Additional research will be needed to specifically test the causes of the SO facet effects. Because of the importance of distinguishing between teacher sorting and co-construction effects, it is useful here to consider what evidence could be helpful in distinguishing between these two sources of differences in observed teaching quality across contexts. The MET study attempted this through randomizing students across classrooms within-schools, which effectively eliminated within-school, between-teacher student sorting as an explanation for the SO facet effects (e.g. Garrett & Steinberg, 2015). Garret and Steinberg (2015), after removing all within-school teacher sorting through randomizing students to teachers, found much of their ability to predict teacher quality (i.e. VA scores) using FFT scores was lost, suggesting co-construction was at play. However, the weak implementation of the randomization process significantly reduced power to detect effects, which may have driven the findings. Further, the difference between within-teacher and between-teacher within-school effects was not broken down. Efforts such as this are an important step in understanding whether co-construction or teacher sorting explains these between-teacher facets, but, at least in the UTQ data, students' prior achievement had a statistically significant effect on the average observed teaching quality between schools (i.e.

schools with higher achieving students had higher observed teaching quality). Thus, student's prior achievement affects observed teaching quality in ways that cannot be examined through within-school experiments. The effects of classroom composition found in this study, however, were within-school/between-teacher effects so the MET randomization data could inform these effects. In general, though, research that observes teachers in multiple school contexts is necessary. This could take the form of observational longitudinal studies where teachers are followed as they move across schools, but the endogenous choice of moving schools will affect the generalizability of these studies. Instead, experiments that incentivize teachers to move schools and then capture the effect on observed teaching quality of this change are necessary to get a true sense of how contexts (especially school context) affect observed teaching quality.

Overall, then, a conclusion from the current study is that SO facets like student composition have a meaningful impact on estimates of teacher quality. Further, under the assumption of co-construction (but not sorting) the use of the SO model is appropriate. The effect of this adjustment is to reduce instrument reliability enough to make it nearly impossible to differentiate teacher quality between teachers with any precision. This is an important finding and one that calls out for more research.

VI.2.4. Is there evidence of instrument bias? The models presented here also showed some evidence of bias across instruments. For most facets, the estimated effect of the facet was consistent across instruments. This means all instruments detected the same shift in observed teaching quality across the levels of the facet, which I interpreted as evidence that observed teaching quality truly changes across the levels of the facet. However, some facets, especially the content domain facets, showed differential effects on teaching quality across instruments, which I interpreted as a sign of instrument bias. I focus here only on instrument bias in the CI facets (and do not further discuss possible bias across live scoring).

The effects of the CI facets are of particular interest when it comes to the question of instrument bias because these facets, I have argued, were the most likely to lead to instrument bias. In fact, there was indication of bias, mostly for the PLATO instrument. The effect of reading lessons, literature lessons, writing lessons, and discussion lessons on PLATO scores was significantly larger than the effect of these lesson types on CLASS scores and FFT scores. Except for reading lessons, differences in the size of the effect across instruments was driven by a large positive effect on PLATO scores and positive, but near-zero effect on CLASS and FFT scores. For reading only, the effect on FFT scores was negative and the effect on PLATO scores was positive. Thus, the evidence for bias is largest for reading, because the reading effect is not dependent on the assumption that the teacher standard deviation metric appropriately scaled the parameter estimates to be equal across instruments. Interpreting these biases is somewhat complex because the same rater provided PLATO scores and the log scores that created the CI facets. Given that rater error across items is known to be correlated (McCaffrey, et al., 2014), correlated rater error could explain this. For example, a rater who rates a lesson as scoring high in use of text in instruction (a PLATO item) may be more likely to rate a lesson as incorporating a reading component (even after controlling for whether a lesson is a reading lesson) than is a rater who does not notice the use of text in instruction. Alternatively, PLATO is designed to measure ELA instruction and so could capture aspects of instruction that are more sensitive to differences across content domain and interaction structure than are the items on CLASS and FFT. An examination of the item-level models in Appendix D showed that these biases were linked to the specific aspects of instruction measured by each instrument. Only some features of instruction vary across the CI facets and, *only* when an instrument measures those features, does it show an effect for the facet. This bias, then, could also be understood as construct under-representation or construct-irrelevant variance in observation instruments. Given only the UTQ data, it is impossible to empirically distinguish these explanations.

There was also evidence of instrument bias across FFT scores and CLASS scores for grammar lessons. Grammar lessons had a marginally positive effect on CLASS scores and a negative effect on FFT scores. This is the strongest evidence of instrument bias, given the independence of scores from the PLATO log rater and the different directions of effects. Again here, I found evidence that this bias was driven by construct under-representation or construct-irrelevant variance with grammar lessons lacking "academic press" (Shouse, 1996) and discussions, which was captured by FFT more than CLASS.

Bias across instruments can play an important role in selecting observation instruments. If we assume that no instrument can fully capture all possible aspects of teaching quality due to limits on possible instrument length and complexity, then when choosing an instrument to use, one would have to select an instrument that captures aspects of teaching quality that are the most important. This, of course, can be done in part by close examination of an instrument, but evidence that shows instruments that do (or do not) respond to preferred methods of instruction can also play an important role in selecting an instrument. For example, if one believes discussions are inherently more effective ways of teaching than lectures/recitation, knowing that an instrument (such as CLASS) rates discussions as higher quality, on average, while other instruments (such as FFT) do not is quite useful. In terms of estimating teacher quality, however, this evidence of bias is more troubling because the biased model should be adjusted for the hidden facets across which bias occurs, but there is no way to tell which of the models is biased when evidence of bias is indicated.

VI.2.5. How well have we explored hidden facet effects? In this section, I reflect on how well hidden facets were explored in this dissertation, dividing the discussion between the CI facets, which vary within-days between-occasions, and the SD and SO facets, which vary between-days and between-teachers. My ability to explore the CI facets was limited by three factors. The first was the use of fixed-time occasions, which I have discussed briefly. I treated CI facets as varying across days, but in reality, they vary across occasions. Lessons do not naturally form 15 minute occasions, and this division creates artificial boundaries that cut across natural divisions in the lesson. If lessons were instead divided into occasions based on naturally formed breaks such as lesson events (Clarke et al., 2007) or occasions with consistent content focus and grouping structure (Carlisle et al., 2011; Stodolsky, 1984), then occasions would more clearly and precisely represent a focus on a specific content domain, instructional grouping, interaction structure, or other factor. This would allow a more precise examination of the effect of CI facets. My approach of identifying lessons with a sustained focus on a content domain or interaction structure is crude compared to an approach of using natural lesson occasions. The crudeness of the occasion-level CI facets in the UTQ data likely led to more error in the creation of facets and lower power to detect the effect of the facet (Williams & Zimmerman, 1989).

A second factor limiting my exploration of CI facets is a lack of days observed for any given teacher. As I argued, the main effects of facets (i.e. average mean difference in observed teaching quality) that I tested for represent the most basic way that a hidden facet might affect teaching quality. Exploring more complicated effects would require estimating different teacher scores for each level of the facet of interest, which in turn requires observing teachers at each level of the hidden facet on multiple days. This would allow me both to more clearly identify the effects of hidden facets on observed teaching quality as acting within-teachers and to examine if the variance of the planned facets of measurement differed across facets. For example, it is possible that lectures have a narrower distribution of teacher variance than small group work. This could happen if all teachers have a relatively high level of skill conducting lectures, due either to more experience lecturing or lectures being inherently easier to conduct than small group work. Estimating separate teacher quality estimates for each level of the facet would allow a more complete exploration of the role of CI facets. Note that this same argument could be applied to SD facets, though I would argue that estimating a teacher's skill in teaching reading and writing separately is more useful than, say, estimating their skill in teaching in the fall and spring separately. This type of analysis is generally not possible for the SO facets because teachers are not observed across the full range of student characteristics or grades (at least for UTQ). The effects of facets on observed teaching quality that I identified in this thesis, then, barely touch on how these facets might affect teaching quality.

The last factor limiting my exploration of CI facets is the limited scope and reliability of the PLATO log. While the PLATO log captured the full range of content domains for English, the interaction structure items were both limited and measured with a great deal of error. Increased measurement error leads to decreased power to detect effects (Williams & Zimmerman, 1989), limiting the ability to truly explore the interaction structure facets. The PLATO log also conflated PLATO scores and CI facets through the common rater providing both sources of information, which limited my ability to explore instrument bias and to accurately estimate the relationship between PLATO scores and CI facets. Additionally, a richer exploration of CI facets would allow for a more complete exploration of how broad classroom processes affect teaching quality. A number of facets previously identified as potentially important or identified as affecting teaching quality include the sequence of content and lessons (Gage & Needels, 1989; Staub 2007; Garrison & Macmillian, 1984), cognitive rigor of lesson (Grossman et al., 2014; Walkington & Marder, 2014), and instructional grouping (Curby et al., 2011; Plank & Condliffe, 2011, 2013). Exploring the effect of commercial curricula on teaching quality would also be interesting.

I was able to better explore the effects of SD and SO facets because they do not suffer from the way occasions were defined in the UTQ study. My analyses showed how both sets of facets affected estimates of teacher quality and explored the level of nesting at which SO facets affected observed teaching quality, demonstrating that prior achievement acted between-schools and student demographics acted within-schools. However, additional work is necessary to explore the generalizability of these effects, especially for teacher evaluation programs in practice and for the level of nesting at which the hidden facets affect observed teaching quality. Moreover, as with CI facets, I was only able to test average mean effects of the SD and SO facets, assuming the effect was constant across teachers and schools, an assumption which should be explored in further work. As I have argued, these constant mean effects are the most simple of possible ways that hidden facets might affect observed teaching quality. Notably though, collecting richer data on each teacher across all levels of a hidden facet is very difficult for the SO facets because teachers are rarely observed across the full range of these facets, much less across multiple schools. Further, my exploration of school effects was lacking in this thesis because the UTQ study did not have a representative sample of teachers from each school, but only included volunteers, leaving me unable to get accurate school means for estimating the impact of schools.

Overall, then, this dissertation has just begun to scratch the surface of understanding how aspects of the lesson observed affect teaching quality. As I argued before, understanding the impact of CI facets can increase the precision of measurement if the effect of the facet is within-teachers and reduce bias when these facets affect between-teacher differences in observed teaching quality. Further, as I will discuss below, they can help guide school and district professional development efforts.

VI.2.6. Should we adjust for hidden facets? One of the most important questions stemming from this dissertation is whether observation scores should be adjusted for the effects of the identified hidden facets. This question, I will argue, has different answers based on the purpose of scores from observation instruments. I speak first about adjusting scores for evaluation systems and next for adjusting scores in research studies. There are some benefits to using the raw mean (i.e. unadjusted scores) of observed teaching quality in evaluation systems. Without adjusting scores, scores from observation instruments can be used as a criterion-referenced measure (J. J. Cohen & Goldhaber, 2016). For example, FFT defines a score of 3 as representing proficient performance. This is beneficial because it holds teachers to an external standard rather than making comparative judgments between teachers. Comparative judgments of teachers may discourage teachers from supporting each other and working together because teachers are judged relative to their peer's performance (J. J. Cohen & Goldhaber, 2016). After making adjustments, the criterion referenced nature of observation scores is muddied (although technically recoverable).

In teacher evaluation systems, observation instruments have both formative feedback and summative feedback purposes. The formative feedback goal of observation requires teachers to get immediate and direct feedback. This feedback will almost certainly be based on the unadjusted scores because an observation instrument's scoring rubric directly links observed classroom behavior to unadjusted scores, allowing the feedback based on unadjusted scores to directly link to specific classroom interactions. Further, adjusting scores takes too much time to provide teachers with immediate feedback on performance. Adjusted scores could be used for the summative purposes of teacher evaluation systems, however, but this will likely result in (at least some) teachers receiving discrepant information from immediate formative feedback on unadjusted scores and later summative feedback on adjusted scores, potentially causing confusion for teachers and damaging trust and confidence in the evaluation system. (cf. Cantrell & Scantlebury, 2011; Kraft & Gilmour, 2016; Bell et al., 2015; Jiang, Sporte, & Luppescu, 2015). Indeed, experience trying to help teachers understand VA scores suggests that explaining statistical adjustments to teachers can be a challenge (Amrein-Beardsley & Collins, 2012; Goldring et al., 2015).

In light of the benefits of using criterion-referenced (unadjusted) scores in evaluation systems, I would argue that there must be clear and substantive benefits of adjusting observation scores to justify the use of adjusted scores in practice. Based on the UTQ data, the effects of the SD and CI facets are too small to justify such adjustments. However, this does not mean that we should ignore these facets. The effects of these facets can still be controlled by randomly sampling of days and by stratifying sampling across time to the extent that is possible. This limits the impact of facets on estimates of teacher quality, which was found in the UTQ data, where estimates of teacher quality across models had surprisingly high correlations. In fact, any recommendations I can make in this vein are conditioned on well-controlled sampling, similar to the UTQ study because any non-ignorable sampling is likely to lead hidden facets to have much larger effects than estimated in this thesis, as I have discussed before.

Thus, research must verify that the SD and CI facets have a minimal effect on scores in specific evaluation systems and on specific observation instruments. Teacher evaluation systems should collect as much data as is feasible about the observation process and lessons being observed in order to explore the impact, in their data, of possible adjustments, making their final decision based on those analyses. Given these analyses replicate the findings in this thesis in the evaluation context, the simple averaging of scores (without adjustments) should be sufficient to estimate a measure of teacher quality.

In fact, understanding the effect of hidden facets on observed teaching quality and estimates of teacher quality can provide a lens into the health of the evaluation system, even if the hidden facets are not being used to formally adjust teacher scores. The CI facets are of particular interest here because they can be used to develop professional development opportunities targeted to the skills of the school or district's teachers. For example, imagine that scores on FFT's culture of learning and engaging students in learning items are low on grammar lessons across a district. The district, upon learning this, might want to develop a professional development series that focuses on how to provide more cognitively engaging and intellectually rigorous grammar instruction, targeting both a content domain that teachers struggle to teach and the specific aspects of instruction that are most difficult to achieve within that content domain. Thus, beyond the measurement question of adjusting scores immediate to this thesis, understanding how CI facets affect observed teaching quality can be beneficial for teacher learning and the design of school improvement programming.

Whether to adjust for the SO facets is a bit more complex. Teacher quality scores from models that adjust for the SO facets are different from those that do not make these adjustments. The difficulty in knowing whether to adjust for the SO facets comes from the need to extrapolate scores across facets. Such extrapolation is unavoidable, unless one decides never to make comparisons across teachers teaching in different contexts. If we use unadjusted scores, then comparing teachers who teach in classrooms with different student characteristics assumes that teacher sorting leads to these differences in observed teaching quality across classrooms with different student characteristics. If we use adjusted scores, on the other hand, we assume co-construction is causing the difference in observed teaching quality across classrooms. In either case, if the assumption is wrong, then comparisons of teachers across levels of the SO facets will be biased. The assumption we make determines which teachers' score estimates might be biased. If we assume teacher sorting but are wrong about this, we inadvertently "punish" teachers teaching disadvantaged students by not properly accounting for how difficult it is to teach these students. If we assume co-construction is present and are wrong, we inadvertently "punish" teachers teaching advantaged students by improperly adjusting away true differences in teacher quality. Given that schools and districts often have trouble filling vacancies at schools serving disadvantaged students, we may want to err on the side of adjusting scores, though some have argued that this unfairly allows lower ability teachers to teach disadvantaged students (assuming teacher sorting is the cause of the difference in observed teaching quality). Until we can distinguish between teacher sorting and co-construction effects, then, it is not clear whether adjustments should be made or not. In practice, then, it seems that not adjusting for facets is the best solution, though concerns about the impact of not adjusting for SO facets might lead some to adjust for these facets. In any case, the difference in teacher quality estimates with and without adjusting for the SO facets is not very large on average (but it is large for some specific teachers).

In contrast, I would recommend that adjustments always be made in research efforts, though estimating teacher quality with and without adjustments will often be the best course. I focus my comments here on research that looks at teacher quality over time to evaluate intervention efforts or to examine teacher growth. Adjusting for the SD facets should be uncontroversial because the timing of when a teacher is observed and who did the observation should play no role in estimating teacher quality. However, the impact of adjusting for these facets on estimation is so small in a well-designed system that it may not be necessary.

The question of adjusting for the CI and SO facets is more complex. The concern here is distinguishing between changes in observed teaching quality that stem from differences in the types of lessons that are observed and from changes in the composition of classrooms from true differences in teacher quality. Unadjusted estimates of teacher quality capture only changes in observed teaching quality, no matter the source. Adjusted estimates of teacher quality capture differences in a teacher's ability to teach within levels of the hidden facets used to adjust scores. This distinction is important. For example, imagine that teachers receive an intervention designed to promote student-centered instruction. This intervention could change observed teaching quality in a number of ways (which are not mutually exclusive). First, teachers could try to be "helpful" by making sure that researchers observe them teaching in student-centered ways. This can result in differences in observed teaching quality due to differences in the way days of instruction are sampled during baseline and post-intervention, which would in fact bias estimates of the impact of the intervention. Second, teachers could adopt some example lessons such that they engage in more student-centered instruction, but do not change how they conduct such instruction (i.e. a shift in frequency and not quality). Third, teachers could begin to engage in higher quality student-centered instruction, where teachers' skill in engaging in such instruction increases (i.e. a shift in quality and not frequency). Unadjusted teacher quality estimates will not be able to distinguish between these three explanations while adjusted teacher quality estimates test for only the third explanation. If unadjusted teacher quality estimates show a gain but adjusted teacher quality does not show a gain, the gain must be caused by one of the first two explanations (which can only be distinguished by careful sampling).

I argue for the use of adjusted estimates of teacher quality in research under the assumption that the third explanation of differences in observed teaching quality is usually the desired target of explanation because it reflects a growth in teacher skill and ability. However, comparing the two estimates is usually the most informative because it would allow, for example, a researcher to conclude that observed teaching quality increased as a result of the intervention to promote student-centered instruction, but this increase was caused solely by an increase in the frequency with which teachers were observed using classroom discussions rather than an increase in the quality of classroom discussions. This sort of conclusion provides a better understanding of how the intervention changed instructional practice than would be possible using only unadjusted or only adjusted teacher quality estimates. The same arguments can be used to argue for comparing teacher quality estimates with and without adjusting for SO facets to distinguish between actual teacher skill development and shifts in the composition of classrooms. Thus, the benefits of adjusting for hidden facets when estimating teacher quality is more clear in research (especially when adjusted estimates of teacher quality are compared with unadjusted estimates) while the pitfalls of accommodating formative and summative feedback do not exist.

VI.3. Concurrent Validity of Teacher Quality Estimates

I began this dissertation by discussing the distinction between teacher quality and observed teaching quality, arguing that teacher quality is the construct of interest. The question naturally arises as to whether one can successfully generalize the observed teaching quality to obtain a true measure of teacher quality. I argued that the size of the correlation between the estimate of teacher quality from a model and the teacher's VA score is a proxy measure for how well I have obtained a true measure of teacher quality. Using the concurrent validity with VA scores, I showed that the estimates of teacher quality from observation instruments did have a significant association with VA scores. However, the different models (i.e. Base model, SD model, CI model, and SO model) produced different estimates of teacher quality, raising the question of whether one model produced estimates that were more valid representations of teacher quality than other models. Validity could differ across model estimates because correcting for instrument bias, which I found for the CI facets, or correcting for bias caused by between-teacher facet effects, which occurred for the CI and SO facets, leads to a better estimate of teacher quality. For example, imagine co-construction explains the effect of the SO facets, estimates of teacher quality from the Base model would incorrectly show large differences in teacher quality across schools and this error would lead to a reduced correlation of teacher quality estimates from the base model to true teacher quality. I tested for differential validity of teacher quality estimates across models, but was unable to find any, possibly due to low power to detect these effects. The low power was the result of high correlations of teacher scores estimates across models which implies that any bias, should it exist, must be small, at least in the UTQ data. Thus, I was not able to provide any evidence regarding which model provided the best estimated of teacher quality.

The most important result of my concurrent validity analysis was not so much the information it provided about correlation of CLASS, FFT, and PLATO scores to VA scores but rather what was learned about the problems of correlated measurement error across these two ways of measuring teaching quality. Measuring teacher quality is a very complex endeavor and estimates from both classroom observation data and student achievement data will have many sources of measurement error. Unfortunately, measurement error involved in using these two sources of data to measure teaching quality will often be correlated, biasing estimates of concurrent validity. In the UTQ data, for example, I found that students' prior-achievement (and the teacher's school) was related both to the previous year's VA scores and to the teacher’s observation scores. Since the students' prior-achievement may be considered a source of error, this implies a shared measurement error. Additionally, using the current year VA score will not overcome this challenge because the same students contribute to the current year VA score and classroom observation score, which likely leads to some bias (Lockwood & McCaffrey, 2012; Lockwood & McCaffrey, 2014).

Thus, for both the prior and current VA scores in the UTQ data, there is, arguably, shared error variance between the VA scores and the estimates of teacher quality based on classroom observation data, and this leads to biased estimates of the validity of estimated teacher quality. Further, this shared error variance will lead to the curious effect that the correlation of VA scores with adjusted observation scores can be lower than the correlation with unadjusted observation scores, where the adjustment eliminates a source of shared error variance, *even when* the adjusted scores are a better measure of teacher quality. This occurs because the shared error variance (i.e. $cov(E_a^{VA}, E_b^{Obs})$) contributes to the covariance of measures (i.e. $cov(X_a^{VA}, X_b^{Obs}) = cov(T + E_a^{VA}, T + E_b^{Obs}) = var(T) + cov(E_a^{VA}, E_b^{Obs})$ where E_b^{Obs} is zero in the adjusted scores because adjustment controls for E_b^{Obs}). This highlights why it is vital to understand any possible sources of correlated errors before interpreting a relationship between measures of the same construct. This has not been done well in past education research, as studies routinely use prior year VA scores to validate observation instruments without controlling for the correlation of each measure with student prior achievement (Blazar et al., 2016; Chaplin et al., 2014; Cohen, 2015a; Cohen & Grossman, 2011; Kane et al., 2013, 2012; Mihaly & McCaffrey, 2014; Milanowski, 2011; Schacter & Thum, 2004; Wayne et al., 2016).

VI.4. Conclusion

Observation instruments are tasked with the challenge of generalizing from a small number of situated measurements of teaching quality in order to capture the teacher-level, stable construct of teacher quality. This usually involves both generalizing scores across hidden facets when a teacher is observed across a range of levels of these facets *and* extrapolating teacher scores across measurement facets when teachers are observed in only a single level of the facet. This is an exceptionally difficult challenge. Any given day of instruction may have an untold number of facets that affect observed teaching quality independent of teacher quality. There is also the threat that instruments themselves are biased such that teaching quality is not measured accurately. Developing a deeper understanding of the hidden measurement facets that contribute to sampling error in estimates of teacher quality, instrument bias, and biases in estimates of teacher quality is vital if we are to interpret estimates of teacher quality as a true measure of teacher quality (i.e. a teacher trait capturing that teachers' general ability to engage in high-quality teaching).

References

Allen, J. P., Gregory, A., Mikami, A., Lun, J., Hamre, B. K., & Pianta, R. C. (2013). Observations of Effective Teacher-Student Interactions in Secondary School Classrooms: Predicting Student Achievement With the Classroom Assessment Scoring System–Secondary. *School Psychology Review*, *42*(1), 76–97. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eft&AN=86877020

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS) in the Houston Independent School District (HISD): Intended and Unintended Consequences. *Education Policy Analysis Archives*, *20*(0), 12. https://doi.org/10.14507/epaa.v20n12.2012

Attali, D. (2016). *Colourpicker: A colour picker tool for shiny and for selecting colours in plots*. Retrieved from https://CRAN.R-project.org/package=colourpicker

Aust, F., & Barth, M. (2016). *Papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An Argument Approach to Observation Protocol Validity. *Educational Assessment*, *17*(2-3), 62–87. https://doi.org/10.1080/10627197.2012.715014

Bell, C. A., Jones, N., Lewis, J., Qi, Y., Kirui, D., Stickler, L., & Liu, S. (2015). *Understanding Consequential Assessment Systems of Teaching: Year 2 Final Report to Los Angeles Unified School District*. ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RM-15-12.pdf

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality: Challenges in observer thinking. In *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project* (pp. 50–97). San Francisco, CA: Jossey-Bass.

Blazar, D., Litke, E., & Barmore, J. (2016). What Does It Mean to Be Ranked a “High” or “Low” Value-Added Teacher? Observing Differences in Instructional Quality Across Districts. *American Educational Research Journal*, *53*(2), 324–359. https://doi.org/10.3102/0002831216630407

Brennan, R. L. (2001). *Generalizability Theory*. New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4757-3456-0

Bretz, R. D., Milkovich, G. T., & Read, W. (1992). The Current State of Performance Appraisal Research and Practice: Concerns, Directions, and Implications. *Journal of Management*, *18*(2), 321–352. https://doi.org/10.1177/014920639201800206

Brophy, J. (2006). Observational Research on Generic Aspects of Classroom Teaching. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 755–780). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Brown, J. L., Jones, S. M., LaRusso, M. D., & Aber, J. L. (2010). Improving Classroom Quality: Teacher Influences and Experimental Impacts of the 4Rs Program. *Journal of Educational Psychology*, *102*(1), 153–167. https://doi.org/10.1037/a0018160

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing Schools for Improvement: Lessons from Chicago*. Chicago ; London: University Of Chicago Press.

Burchinal, M., Vandergrift, N., Pianta, R. C., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*, *25*(2), 166–176. https://doi.org/10.1016/j.ecresq.2009.10.004

Bürkner, P.-C. (in press). brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software*.

Cadima, J., Leal, T., & Burchinal, M. (2010). The quality of teacher student interactions: Associations with first graders' academic and behavioral outcomes. *Journal of School Psychology*, *48*(6), 457–482. https://doi.org/10.1016/j.jsp.2010.09.001

Calkins, D., Borich, G. D., Pascone, M., Kugle, C. L., & Marston, P. T. (1977). Generalizability of Teacher Behaviors Across Classroom Observation Systems. *The Journal of Classroom Interaction*, *13*(1), 9–22.

Camburn, E., & Barnes, C. A. (2004). Assessing the validity of a language arts instruction log through triangulation. *The Elementary School Journal*, 49–73.

Cantrell, S., & Scantlebury, J. (2011). *Effective Teaching: What Is It and How Is It Measured?* (VUE). Annenberg Institute for School Reform.

Carlisle, J., Kelcey, B., Berebitsky, D., & Phelps, G. (2011). Embracing the Complexity of Instruction: A Study of the Effects of Teachers' Instruction on Students' Reading Comprehension. *Scientific Studies of Reading*, *15*(5), 409–439. https://doi.org/10.1080/10888438.2010.497521

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, 0013164414539163. Retrieved from http://epm.sagepub.com/content/early/2014/07/03/0013164414539163.abstract

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. https://doi.org/10.1177/0013164413486987

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly*, *27*(3), 529–542. https://doi.org/10.1016/j.ecresq.2011.12.006

Cassidy, D. J., Hestenes, L. L., Hegde, A., Hestenes, S., & Mims, S. (2005). Measurement of quality in preschool child care classrooms: An exploratory and confirmatory factor analysis of the early childhood environment rating scale-revised. *Early Childhood Research Quarterly*, *20*(3), 345–360. https://doi.org/10.1016/j.ecresq.2005.07.005

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional Practice, Student Surveys, and Value-Added: Multiple Measures of Teacher Effectiveness in the Pittsburgh Public Schools* (No. REL 2014-024). Washington, DC: Regional Educational Laboratory Mid-Atlantic. Retrieved from http://eric.ed.gov/?id=ED545232

Clarke, D., Mesiti, C., O'Keefe, C., Xu, L. H., Jablonka, E., Mok, I. A. C., & Shimizu, Y. (2007). Addressing the challenge of legitimate international comparisons of classroom practice. *International Journal of Educational Research*, *46*(5), 280–293. https://doi.org/10.1016/j.ijer.2007.10.009

Cohen, D. K., & Ball, D. L. (1999). *Instruction, Capacity, and Improvement* (CPRE research report series No. RR-43). Consortium for Policy Research in Education.

Cohen, J. J. (2015a). Challenges in Identifying High-Leverage Practices. *Teachers College Record*, *117*(7).

Cohen, J. J. (2015b). Explicit Instruction Across Elementary Math and Language Arts. In. Presented at the AERA, Chicago, IL.

Cohen, J. J., & Brown, M. (2016). Teaching Quality Across School Settings. *The New Educator*, *12*(2), 191–218. https://doi.org/10.1080/1547688X.2016.1156459

Cohen, J. J., & Goldhaber, D. (2016). Building a More Complete Understanding of Teacher Evaluation Using Classroom Observations. *Educational Researcher*, *45*(6), 378–387. https://doi.org/10.3102/0013189X16659442

Cohen, J. J., & Grossman, P. (2011). Of Cabbages and Kings: Classroom Observations & Value-Added Measures. In. Presented at the Annual meeting of AERA. Retrieved from http://platorubric.stanford.edu/2011%20AERA%20paper%20Cabbages%20%20Kings.pdf

Cohen, J. J., & Grossman, P. (2016). Respecting complexity in measures of teaching: Keeping students and schools in focus. *Teaching and Teacher Education*, *55*, 308–317. https://doi.org/10.1016/j.tate.2016.01.017

Connor, C. M., Morrison, F. J., & Petrella, J. N. (2004). Effective Reading Comprehension Instruction: Examining Child x Instruction Interactions. *Journal of Educational Psychology*, *96*(4), 682–698. https://doi.org/10.1037/0022-0663.96.4.682

Connor, C. M., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., Schatschneider, C. (2009a). The ISI Classroom Observation System: Examining the Literacy Instruction Provided to Individual Students. *Educational Researcher*, *38*(2), 85–99. https://doi.org/10.3102/0013189X09332373

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., … Morrison, F. J. (2009b). Individualizing student instruction precisely: Effects of Child x Instruction interactions on first Graders' literacy development. *Child Development*, *80*(1), 77–100. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8624.2008.01247.x/full

Cor, M. K. (2011). Investigating the Reliability of Classroom Observation Protocols: The Case of PLATO. Retrieved from http://platorubric.stanford.edu/Cor%20M%20K%20%20(2011).pdf

Cortina, K. S., Miller, K. F., McKenzie, R., & Epstein, A. (2015). Where Low and High Inference Data Converge: Validation of CLASS Assessment of Mathematics Instruction Using Mobile Eye Tracking with Expert and Novice Teachers. *International Journal of Science and Mathematics Education*, *13*(2), 389–403. https://doi.org/10.1007/s10763-014-9610-5

Croninger, R. G., & Valli, L. (2009). "Where Is the Action?" Challenges to Studying the Teaching of Reading in Elementary Classrooms. *Educational Researcher*, *38*(2), 100–108. https://doi.org/10.3102/0013189X09333206

Curby, T. W., LoCasale-Crouch, J., Konold, T. R., Pianta, R. C., Howes, C., Burchinal, M., … Barbarin, O. (2009). The Relations of Observed Pre-K Classroom Quality Profiles to Children's Achievement and Social Competence. *Early Education and Development*, *20*(2), 346–372. https://doi.org/10.1080/10409280802581284

Curby, T. W., Rimm-Kaufman, S. E., & Ponitz, C. C. (2009). Teacher-child interactions and children's achievement trajectories across kindergarten and first grade. *Journal of Educational Psychology*, *101*(4), 912–925. https://doi.org/10.1037/a0016647

Curby, T. W., Rudasill, K. M., Edwards, T., & Pérez-Edgar, K. (2011). The role of classroom quality in ameliorating the academic and social risks associated with difficult temperament. *School Psychology Quarterly*, *26*(2), 175–188. https://doi.org/10.1037/a0023042

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

Dahl, D. B. (2016). *Xtable: Export tables to latex or html*. Retrieved from https://CRAN.R-project.org/package=xtable

Danielson, C. (2000). *Teacher Evaluation to Enhance Professional Practice*. Association for Supervision & Curriculum Development.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press. Retrieved from http://statwww.epfl.ch/davison/BMA/

Dee, T. S., & Wyckoff, J. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297. https://doi.org/10.1002/pam.21818

Deng, N., & Hambleton, R. K. (2013). Evaluating CTT- and IRT-Based Single-Administration Estimates of Classification Consistency and Accuracy. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New Developments in Quantitative Psychology* (pp. 235–250). Springer New York. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4614-9348-8_15

Dowle, M., & Srinivasan, A. (2016). *Data.table: Extension of 'data.frame'*. Retrieved from https://CRAN.R-project.org/package=data.table

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

Ferris, G. R., Munyon, T. P., Basik, K., & Buckley, M. R. (2008). The performance evaluation context: Social, emotional, cognitive, political, and relationship components. *Human Resource Management Review*, *18*(3), 146–163. https://doi.org/10.1016/j.hrmr.2008.07.006

Floman, J. L., Hagelskamp, C., Brackett, M. A., & Rivers, S. E. (2016). Emotional Bias in Classroom Observations Within-Rater Positive Emotion Predicts Favorable Assessments of Classroom Quality. *Journal of Psychoeducational Assessment*. https://doi.org/10.1177/0734282916629595

Fox, J. (2016). *Polycor: Polychoric and polyserial correlations*. Retrieved from https://CRAN.R-project.org/package=polycor

Gage, N. L., & Needels, M. C. (1989). Process-Product Research on Teaching: A Review of Criticisms. *The Elementary School Journal*, *89*(3), 253–300.

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

Garrison, J. W., & Macmillian, C. J. B. (1984). A Philosophical Critique of the Process-Product Research on Teaching. *Educational Theory*, *34*(3).

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). *The content, predictive power, and potential bias in five widely used teacher observation instruments* (Indexes; Offices No. REL 2017-191). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs

Gitomer, D. H., & Bell, C. A. (2013). Evaluating Teaching and Teachers. In *APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education*. American Psychological Association.

Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6).

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make Room Value Added Principals' Human Capital Decisions and the Emergence of Teacher Observation Data. *Educational Researcher*, *44*(2), 96–104. Retrieved from http://edr.sagepub.com/content/44/2/96.short

Golman, R., & Bhatia, S. (2012). Performance evaluation inflation and compression. *Accounting, Organizations and Society*, *37*(8), 534–543. https://doi.org/10.1016/j.aos.2012.09.001

Good, T. L. (1979). Teacher Effectiveness in the Elementary school. *Journal of Teacher Education*, *30*(2), 52–64. https://doi.org/10.1177/002248717903000220

Graham, M., Milanowski, A. T., & Miller, J. (2012). *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings* (No. ED532068). Center for Educator Compensation Reform. Retrieved from http://eric.ed.gov/?id=ED532068

Greenacre, M. (2005). *From Correspondence Analysis to Multiple and Joint Correspondence Analysis* (Economics working papers). BBVA Foundation. Retrieved from http://econpapers.repec.org/paper/upfupfgen/883.htm

Grissom, J. A., & Loeb, S. (2016). Assessing Principals' Assessments: Subjective Evaluations of Teacher Effectiveness in Low- and High-Stakes Environments. *Education Finance and Policy*, 1–53. https://doi.org/10.1162/EDFP_a_00210

Grossman, P., Cohen, J. J., & Brown, L. (2014). Understanding Instructional Quality in English Language Arts: Variations in PLATO Scores by Content and Context. In *Designing teacher evaluation systems: New guidance from the measures of effecting project* (pp. 303–331). San Francisco, CA: Jossey-Bass.

Grossman, P., Cohen, J. J., Ronfeldt, M., & Brown, L. (2014). The Test Matters The Relationship Between Classroom Observation Scores and Teacher Value Added on Multiple Types of Assessment. *Educational Researcher*, *43*(6), 293–303. https://doi.org/10.3102/0013189X14544542

Grossman, P., Loeb, S., Cohen, J. J., & Wyckoff, J. (2013). Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores. *American Journal of Education*, *119*(3), 445–470. https://doi.org/10.1086/669901

Grossman, P., Loeb, S., Cohen, J. J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for Measure: The Relationship Between Measures of Instructional Practice in Middle School English Language Arts and Teacher's Value Added Scores* (No. Working Paper 45). NBER WORKING PAPER SERIES. Retrieved from http://www.nber.org/papers/w16015

Gwet, K. L. (2012). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters, 3rd Edition* (3rd edition). Gaithersburg, MD: Advanced Analytics, LLC.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first grade classroom make a difference for children at risk of school failure? *Child Development*, *76*(5), 949–967.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A., Jones, S. M., … Hamagami, A. (2013). Teaching through Interactions: Testing a Developmental Framework of Teacher Effectiveness in over 4,000 Classrooms. *The Elementary School Journal*, *113*(4), 461–487. https://doi.org/10.1086/669616

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An Examination of Rater Drift Within a Generalizability Theory Framework. *Journal of Educational Measurement*, *46*(1), 43–58. https://doi.org/10.1111/j.1745-3984.2009.01068.x

Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and ltems. *Applied Psychological Measurement*, *9*(2), 139–164. https://doi.org/10.1177/014662168500900204

Hill, H. C., Ball, D. L., Blunk, M., Goffney, I. M., & Rowan, B. (2007). Validating the Ecological Assumption: The Relationship of Measure Scores to Classroom Teaching and Student Learning. *Measurement: Interdisciplinary Research and Perspectives*, *5*(2-3), 107–118. https://doi.org/10.1080/15366360701487138

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012a). When Rater Reliability Is Not Enough Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*, *41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., … Lynch, K. (2012b). Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation. *Educational Assessment*, *17*(2-3), 88–106. https://doi.org/10.1080/10627197.2012.715019

Ho, A. D., & Kane, T. J. (2013). The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project. *Bill & Melinda Gates Foundation*. Retrieved from http://eric.ed.gov/?id=ED540957

Hoffman, J. V., Sailors, M., Duffy, G. R., & Beretvas, S. N. (2004). The Effective Elementary Classroom Literacy Environment: Examining the Validity of the TEX-IN3 Observation System. *Journal of Literacy Research*, *36*(3), 303–334. https://doi.org/10.1207/s15548430jlr3603_3

Holtzapple, E. (2003). Criterion-Related Validity Evidence for a Standards-Based Teacher Evaluation System. *Journal of Personnel Evaluation in Education*, *17*(3), 207–219. https://doi.org/10.1007/s11092-005-2980-z

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*(4), 403–424. https://doi.org/10.1037/1082-989X.4.4.403

Jiang, J. Y., & Sporte, S. E. (2016). *Teacher Evaluation in Chicago Differences in Observation and Value- Added Scores by Teacher, Student, and School Characteristics*. UChicago Consortium on School Research.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher Perspectives on Evaluation Reform Chicago's REACH Students. *Educational Researcher*, *44*(2), 105–116. Retrieved from http://edr.sagepub.com/content/44/2/105.short

Joe, J. N., McClellan, C., & Holtzman, S. L. (2014). Scoring Design Decisions: Reliability and the Length and Focus of Classroom Observations. In *Designing teacher evaluation systems: New guidance from the measures of effecting project* (pp. 415–443). San Francisco, CA: Jossey-Bass.

Joe, J. N., Tocci, C. M., Holtzman, S. L., & Williams, J. C. (2013). Foundations of Observation. Retrieved from http://www.gtlcenter.org/sites/default/files/MET-ETS_Foundations_of_Observation.pdf

Kane, T. J., & Cantrell, S. (2010). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project: Research Paper*. MET Project Research Paper, Bill & Melinda Gates Foundation.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project. *Bill & Melinda Gates Foundation*. Retrieved from http://eric.ed.gov/?id=ED540959

Kane, T. J., Staiger, D. O., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., & Parker, D. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*. Seattle, WA: Bill & Melinda Gates Foundation, Measures of Effective Teaching Project. Retrieved from http://eric.ed.gov/?id=ED540960

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources*, *46*(3), 587–613. https://doi.org/10.1353/jhr.2011.0010

Kelcey, B., & Carlisle, J. (2013). Learning About Teachers' Literacy Instruction From Classroom Observations. *Reading Research Quarterly*, *48*(3), 301–317. https://doi.org/10.1002/rrq.51

Kennedy, M. (2010). Approaches to Annual Performance Assessment. In *Teacher assessment and the quest for teacher quality: A handbook*. John Wiley & Sons.

Kleiner, A., Talwalkar, A., Agarwal, S., Stoica, I., & Jordan, M. I. (2013). A general bootstrap performance diagnostic. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 419–427). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2487650

Kraft, M. A., & Gilmour, A. F. (2016). Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness.

Ladson-Billings, G. (2008). Opportunity to Teach: Teacher Quality in Context. In *Measurement Issues and Assessment for Teaching Quality*. Thousand Oaks: SAGE Publications, Inc.

Lazarev, V., & Newman, D. (2015). *How Teacher Evaluation Is Affected by Class Characteristics: Are Observations Biased?* (SSRN Scholarly Paper No. ID 2574897). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2574897

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33. https://doi.org/10.18637/jss.v069.i01

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Lockwood, J. R., & McCaffrey, D. (2012). Reducing Bias in Teacher Value-Added Estimates by Accounting for Test Measurement Error. In. Presented at the SREE.

Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for Test Score Measurement Error in ANCOVA Models for Estimating Treatment Effects. *Journal of Educational and Behavioral Statistics*, *39*(1), 22–52. https://doi.org/10.3102/1076998613509405

Lynch, K., Chin, M., & Blazar, D. (2015). Relationship between observations of elementary teacher mathematics instruction and student achievement: Exploring variability across districts. *Cambridge, MA: National Center for Teacher Effectiveness, Harvard University*. Retrieved from http://scholar.harvard.edu/files/david_blazar/files/lynch_chin_and_blazar_classroom_observations_and_achievement_across_districts_working_paper.pdf

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

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50-60.

Mashburn, A., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2013). Improving the Power of an Efficacy Study of a Social and Emotional Learning Program: Application of Generalizability Theory to the Measurement of Classroom-Level Outcomes. *Prevention Science*, *15*(2), 146–155. https://doi.org/10.1007/s11121-012-0357-3

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward Measuring Instructional Interactions “At-Scale”. *Educational Assessment*, *13*(4), 267–300. https://doi.org/10.1080/10627190802602541

Matsumura, L. C., Slater, S. C., Junker, B., Peterson, M., Boston, M., Steele, M., & Resnick, L. (2006). *Measuring Reading Comprehension and Mathematics Instruction in Urban Middle Schools: A Pilot Study of the Instructional Quality Assessment. CSE Technical Report 681*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from http://eric.ed.gov/?id=ED492885

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2014). Uncovering Multivariate Structure in Classroom Observations in the Presence of Rater Errors. *Educational Measurement: Issues and Practice*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/emip.12061/full

Mihaly, K., & McCaffrey, D. F. (2014). Grade level variation in observational measures of teacher effectiveness. In *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project. New York: John Wiley & Sons*. San Francisco, CA: Jossey-Bass.

Milanowski, A. T. (2011). *Validity Research on Teacher Evaluation Systems Based on the Framework for Teaching*. Retrieved from http://eric.ed.gov/?id=ED520519

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation*, *12*(1), 53–74. https://doi.org/10.1080/13803610500392236

Murphy, K. R., & Deshon, R. (2000). Interrater Correlations Do Not Estimate the Reliability of Job Performance Ratings. *Personnel Psychology*, *53*(4), 873–900. https://doi.org/10.1111/j.1744-6570.2000.tb02421.x

Myford, C. M., & Wolfe, E. W. (2009). Monitoring Rater Performance Over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use. *Journal of Educational Measurement*, *46*(4), 371–389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

Newton, X. A. (2010). Developing indicators of classroom practice to evaluate the impact of district mathematics reform initiative: A generalizability analysis. *Studies in Educational Evaluation*, *36*(12), 1–13. https://doi.org/10.1016/j.stueduc.2010.10.002

Park, Y. S., Holtzman, S., & Chen, J. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. In *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*. San Francisco, CA: Jossey-Bass.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity. *Educational Researcher*, *38*(2), 109–119. https://doi.org/10.3102/0013189X09332374

Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S. L., & La Paro, K. M. (2007). *CLASS-Secondary Manual*. Charlottesville, VA: Teachstone.

Plank, S. B., & Condliffe, B. (2011). Pressures of the Season: A Descriptive Look at Classroom Quality in Second and Third Grade Classrooms. *Baltimore Education Research Consortium*. Retrieved from http://eric.ed.gov/?id=ED535779

Plank, S. B., & Condliffe, B. (2013). Pressures of the Season: An Examination of Classroom Quality and High-Stakes Accountability. *American Educational Research Journal*, *50*(5), 1152–1182. https://doi.org/10.3102/0002831213500691

Polikoff, M. S. (2015). The Stability of Observational and Student Survey Measures of Teaching Effectiveness. *American Journal of Education*, *121*(2), 183–212. https://doi.org/10.1086/679390

Polikoff, M. S., & Porter, A. C. (2014). Instructional Alignment as a Measure of Teaching Quality. *Educational Evaluation and Policy Analysis*, *36*(4), 399–416. https://doi.org/10.3102/0162373714531851

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, *22*(6), 387–400. https://doi.org/10.1016/j.learninstruc.2012.03.002

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*(5), 959–981. https://doi.org/10.1037/0021-9010.93.5.959

R Core Team. (2016a). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

R Core Team. (2016b). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raudenbush, S. W. (2013). What do we know about using value-added to compare teachers who work in different schools? *Carnegie Foundation for the Advancement of Teaching: Carnegie Knowledge Network Knowledge Brief. Accessed March*, *25*, 2015. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/comparing-teaching/

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd edition). Thousand Oaks: SAGE Publications, Inc.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of Value-Added Models for Estimating School Effects. *Education Finance and Policy*, *4*(4), 492–519. https://doi.org/10.1162/edfp.2009.4.4.492

Revelle, W. (2016). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Rimm-Kaufman, S. E., Early, D. M., Cox, M. J., Saluja, G., Pianta, R. C., Bradley, R. H., & Payne, C. (2002). Early behavioral attributes and teachers' sensitivity as predictors of competent behavior in the kindergarten classroom. *Journal of Applied Developmental Psychology*, *23*(4), 451–470. https://doi.org/10.1016/S0193-3973(02)00128-4

Rimm-Kaufman, S. E., Paro, K. M. L., Downer, J. T., & Pianta, R. C. (2005). The Contribution of Classroom Setting and Quality of Instruction to Children's Behavior in Kindergarten Classrooms. *The Elementary School Journal*, *105*(4), 377–394. https://doi.org/10.1086/429948

Rowan, B., & Correnti, R. (2009). Studying Reading Instruction With Teacher Logs: Lessons From the Study of Instructional Improvement. *Educational Researcher*, *38*(2), 120–131. https://doi.org/10.3102/0013189X09332375

Rowan, B., Camburn, E., & Correnti, R. (2004). Using Teacher Logs to Measure the Enacted Curriculum: A Study of Literacy Teaching in Third-Grade Classrooms. *The Elementary School Journal*, *105*(1), 75–101. https://doi.org/10.1086/428803

Rowan, B., Schilling, S. G., Spain, A., Bhandari, P., Berger, D., & Graves, J. (2013). *Promoting High Quality Teacher Evaluations in Michigan: Lessons from a Pilot of Educator Effectiveness Tools*. Institute of Social Research; The University of Michigan.

Sartain, L., Stoelinga, S. R., & Brown, E. (2009). Evaluation of the Excellence in Teaching Pilot: Year 1 Report to the Joyce Foundation. Consortium of Chicago School Research.

Schacter, J., & Thum, Y. M. (2004). Paying for high- and low-quality teaching. *Economics of Education Review*, *23*(4), 411–430. https://doi.org/10.1016/j.econedurev.2003.08.002

Schutz, A., & Moss, P. A. (2004). Reasonable Decisions in Portfolio Assessment: Evaluating Complex Evidence of Teaching. *Education Policy Analysis Archives*, *12*(33). Retrieved from http://eric.ed.gov.proxy.lib.umich.edu/?id=EJ852315

Seidel, T., & Shavelson, R. J. (2007). Teaching Effectiveness Research in the Past Decade: The Role of Theory and Research Design in Disentangling Meta-Analysis Results. *Review of Educational Research*, *77*(4), 454–499. https://doi.org/10.3102/0034654307310317

Shavelson, R. J., & Dempsey-Atwood, N. (1976). Generalizability of Measures of Teaching Behavior. *Review of Educational Research*, *46*(4), 553–611. https://doi.org/10.3102/00346543046004553

Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. *Handbook of Research on Teaching*, *3*, 50–91.

Shouse, R. C. (1996). Academic press and sense of community: Conflict, congruence, and implications for student achievement. *Social Psychology of Education*, *1*(1), 47–68. Retrieved from http://www.springerlink.com/index/10.1007/BF02333405

Shumate, S. R., Surles, J., Johnson, R. L., & Penny, J. (2007). The Effects of the Number of Scale Points and Non-Normality on the Generalizability Coefficient: A Monte Carlo Study.

Applied Measurement in Education, *20*(4), 357–376. https://doi.org/10.1080/08957340701429645

Staub, F. C. (2007). Mathematics classroom cultures: Methodological and theoretical issues. *International Journal of Educational Research*, *46*(5), 319–326. https://doi.org/10.1016/j.ijer.2007.10.007

Steinberg, M. P., & Donaldson, M. L. (2015). The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era. *Education Finance and Policy*, 1–40. https://doi.org/10.1162/EDFP_a_00186

Steinberg, M. P., & Garrett, R. (2016). Classroom Composition and Measured Teacher Performance: What Do Teacher Observation Scores Really Measure? *Educational Evaluation and Policy Analysis*, 1–25. https://doi.org/10.3102/0162373715616249

Steinberg, M. P., & Sartain, L. (2015). Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*, *10*(4), 535–572. https://doi.org/10.1162/EDFP_a_00173

Stodolsky, S. S. (1984). Teacher Evaluation: The Limits of Looking. *Educational Researcher*, *13*(9), 11–18. https://doi.org/10.2307/1174874

Stuhlman, M. W., & Pianta, R. C. (2009). Profiles of Educational Quality in First Grade. *The Elementary School Journal*, *109*(4), 323–342. https://doi.org/10.1086/593936

Sumer, H. C., & Knight, P. A. (1996). Assimilation and contrast effects in performance ratings: Effects of rating the previous performance on rating subsequent performance. *Journal of Applied Psychology*, *81*(4), 436–442.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Walkington, C., & Marder, M. (2014). Classroom Observation and Value-Added Models Give Complementary Information About Quality of Mathematics Teaching. In *Designing teacher evaluation systems: New guidance from the measures of effecting project* (pp. 234–277). San Francisco, CA: Jossey-Bass.

Wang, X. M., Wong, K. F. E., & Kwong, J. Y. Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology*, *95*(3), 546–561. https://doi.org/10.1037/a0018866

Warnes, G. R., Bolker, B., Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., … others. (2015). *Gdata: Various r programming tools for data manipulation*. Retrieved from https://CRAN.R-project.org/package=gdata

Wayne, A. J., Garet, M. S., Brown, S., Rickles, J., Song, M., Manzeske, D., & Ali, M. (2016). *Early Implementation Findings From a Study of Teacher and Principal Performance Measurement and Feedback: Year 1 Report* (No. NCEE 2017-4004). Washington D.C.: IES National Center for Education Evaluation and Regional Assistance.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In *Handbook of Statistics* (Vol. 26, pp. 81–124). Elsevier B.V.

Whitehurst, G. J. R., Chingos, M. M., & Lindquist, K. M. (2014). Evaluating Teachers with Classroom Observations. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2014/05/13-teacher-evaluation/evaluating-teachers-with-classroom-observations.pdf

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Williams, R. H., & Zimmerman, D. W. (1989). Statistical Power Analysis and Reliability of Measurement. *The Journal of General Psychology*, *116*(4), 359–369. https://doi.org/10.1080/00221309.1989.9921123

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from http://yihui.name/knitr/

Appendices

Appendix A – Comparison of Model RE with CI

Table A.1: Variance of the Error Facets with Confidence Interval for the CLASS Models

	Base Model		SD Model		CI Model		SO Model	
Facet	Value	Percent	Value	Percent	Value	Percent	Value	Percent
Teacher ($var(v_t)$)	0.076 (0.054-0.102)	7% (5-9.1)	0.066 (0.046-0.087)	6.4% (4.5-8.3)	0.06 (0.042-0.078)	5.8% (4.1-7.6)	0.031 (0.018-0.045)	3.1% (1.8-4.6)
Day ($var(v_{d:s:t})$)	0.013 (0-0.035)	1.2% (0-3.2)	0.007 (0-0.027)	0.7% (0-2.6)	0.007 (0-0.027)	0.7% (0-2.6)	0.005 (0-0.024)	0.5% (0-2.4)
Occasion ($var(v_{o:d:s:t})$)	0.053 (0.048-0.058)	4.9% (4.4-5.5)	0.052 (0.047-0.056)	5% (4.4-5.5)	0.052 (0.047-0.057)	5% (4.5-5.6)	0.052 (0.047-0.056)	5.2% (4.6-5.7)
Rater ($var(v_r)$)	0.04 (0-0.102)	3.7% (0-8.9)	0.021 (0-0.065)	2.1% (0-6.1)	0.022 (0-0.063)	2.1% (0-6.1)	0.022 (0-0.071)	2.2% (0-7)
Rater-by-Teacher ($var(v_{rt})$)	0 (0-0.036)	0% (0-3.3)	0.012 (0-0.044)	1.1% (0-4.2)	0.013 (0-0.042)	1.3% (0-4.1)	0.014 (0-0.045)	1.4% (0-4.5)
Rater-by-Day ($var(v_{r(d:s:t)})$)	0.141 (0.101-0.158)	13% (9.3-14.6)	0.116 (0.08-0.139)	11.2% (7.7-13.6)	0.114 (0.078-0.136)	11.1% (7.7-13.4)	0.112 (0.08-0.137)	11.3% (8-13.8)
Item-by-Rater ($var(v_{ir})$)	0.225 (0.168-0.289)	20.7% (16.1-25.5)	0.225 (0.168-0.286)	21.8% (17-26.4)	0.225 (0.167-0.287)	21.9% (17-26.7)	0.225 (0.169-0.288)	22.6% (17.8-27.1)
Item-by-Teacher ($var(v_{i(t)})$)	0.029 (0.024-0.034)	2.7% (2.2-3.2)	0.029 (0.024-0.034)	2.8% (2.3-3.4)	0.029 (0.024-0.035)	2.8% (2.3-3.4)	0.029 (0.024-0.035)	2.9% (2.4-3.5)
Item-by-Day ($var(v_{i(d:s:t)})$)	0.128 (0.121-0.136)	11.8% (10.8-12.9)	0.129 (0.122-0.137)	12.5% (11.4-13.6)	0.129 (0.122-0.137)	12.6% (11.6-13.7)	0.129 (0.122-0.137)	12.9% (11.8-14.1)
Item-by-Occasion ($var(v_{i(o:d:s:t)})$)	0 (0-0.008)	0% (0-0.7)	0 (0-0.009)	0% (0-0.9)	0 (0-0.01)	0% (0-0.9)	0 (0-0.008)	0% (0-0.8)
Residual ($var(\epsilon_{ir(o:d:s:t)})$)	0.381 (0.371-0.386)	35.1% (32.3-37.5)	0.377 (0.367-0.382)	36.5% (33.9-38.6)	0.377 (0.367-0.382)	36.7% (34.1-39)	0.377 (0.367-0.381)	37.8% (34.8-40.2)

Note. Each pair of columns shows a separate model. For each regression model, the left column displays the estimated variance and the right column displays the percentage of variance for each error facet.

Table A.2: Variance of the Error Facets with Confidence Interval for the FFT Models

	Base Model		SD Model		CI Model		SO Model	
Facet	Value	Percent	Value	Percent	Value	Percent	Value	Percent
Teacher $(var(v_t))$	0.029	10.7%	0.026	9.7%	0.023	8.9%	0.012	4.7%
	(0.021-0.038)	(7.9-13.7)	(0.019-0.034)	(7.2-12.3)	(0.017-0.03)	(6.5-11.4)	(0.007-0.016)	(2.9-6.5)
Day $(var(v_{d:s:t}))$	0.008	3%	0.008	3%	0.008	2.9%	0.007	2.7%
	(0.001-0.016)	(0.4-5.9)	(0.001-0.015)	(0.4-5.5)	(0.001-0.015)	(0.3-5.7)	(0-0.014)	(0-5.3)
Rater $(var(v_r))$	0.011	4.2%	0.011	4.1%	0.01	3.9%	0.011	4.3%
	(0.003-0.024)	(1-8.5)	(0.003-0.023)	(1-8.3)	(0.002-0.024)	(0.8-8.6)	(0.002-0.023)	(1-8.7)
Rater-by-Teacher	0.005	1.8%	0.005	1.9%	0.005	2%	0	0%
$(var(v_{rt}))$	(0-0.019)	(0-6.6)	(0-0.017)	(0-6.3)	(0-0.017)	(0-6.6)	(0-0.011)	(0-4.6)
Rater-by-Day	0.044	16%	0.038	14.6%	0.038	14.7%	0.043	17.5%
$(var(v_{r(d:s:t)}))$	(0.029-0.054)	(10.5-19.6)	(0.026-0.048)	(9.8-18.4)	(0.025-0.048)	(9.6-18.4)	(0.031-0.05)	(12.3-20.1)
Item-by-Rater	0.011	4%	0.011	4.2%	0.011	4.2%	0.011	4.4%
$(var(v_{ir}))$	(0.008-0.015)	(2.9-5.2)	(0.008-0.015)	(3-5.5)	(0.008-0.014)	(3-5.5)	(0.008-0.015)	(3.2-5.9)
Item-by-Teacher	0.008	2.9%	0.008	2.9%	0.008	3%	0.008	3.1%
$(var(v_{i(t)}))$	(0.006-0.01)	(2.1-3.7)	(0.005-0.01)	(2-3.9)	(0.005-0.01)	(2.1-3.8)	(0.006-0.01)	(2.2-4)
Item-by-Day	0.017	6%	0.017	6.3%	0.017	6.4%	0.017	6.7%
$(var(v_{i(d:s:t)}))$	(0.012-0.021)	(4.5-7.7)	(0.012-0.022)	(4.7-8.2)	(0.013-0.022)	(4.9-8.3)	(0.013-0.022)	(5.1-8.8)
Residual	0.14	51.4%	0.14	53.3%	0.14	53.9%	0.14	56.5%
$(var(\epsilon_{ir(o:d:s:t)}))$	(0.135-0.145)	(48.1-54.3)	(0.135-0.145)	(49.8-56)	(0.135-0.145)	(50.8-56.6)	(0.135-0.145)	(53-59.4)

Note. Each pair of columns shows a separate model. For each regression model, the left column displays the estimated variance and the right column displays the percentage of variance for each error facet.

Table A.3: Variance of the Error Facets with Confidence Interval for the PLATO Models

	Base Model		SD Model		CI Model		SO Model	
Facet	Value	Percent	Value	Percent	Value	Percent	Value	Percent
Teacher ($var(v_t)$)	0.012 (0.007-0.015)	2.8% (1.8-3.7)	0.01 (0.006-0.013)	2.4% (1.5-3.3)	0.007 (0.004-0.01)	1.8% (1.1-2.5)	0.005 (0.002-0.007)	1.3% (0.6-1.9)
Day ($var(v_{d:s:t})$)	0.003 (0-0.008)	0.8% (0-1.8)	0.004 (0-0.008)	0.9% (0-2)	0 (0-0.004)	0% (0-1.1)	0 (0-0.004)	0% (0-1)
Occasion ($var(v_{o:d:s:t})$)	0.017 (0.015-0.019)	4.1% (3.6-4.6)	0.016 (0.014-0.018)	4% (3.5-4.5)	0.016 (0.014-0.018)	4.1% (3.6-4.5)	0.016 (0.014-0.018)	4.1% (3.6-4.6)
Rater ($var(v_r)$)	0.002 (0-0.009)	0.5% (0-2.2)	0 (0-0.003)	0% (0-0.7)	0 (0-0.003)	0% (0-0.8)	0 (0-0.004)	0.1% (0-1)
Rater-by-Teacher ($var(v_{rt})$)	0 (0-0.005)	0.1% (0-1.2)	0 (0-0.005)	0% (0-1.2)	0 (0-0.005)	0% (0-1.2)	0 (0-0.004)	0% (0-1.1)
Rater-by-Day ($var(v_{r(d:s:t)})$)	0.02 (0.014-0.024)	4.8% (3.5-5.7)	0.019 (0.013-0.022)	4.7% (3.3-5.4)	0.018 (0.013-0.021)	4.7% (3.2-5.2)	0.018 (0.013-0.02)	4.7% (3.3-5.2)
Item-by-Rater ($var(v_{ir})$)	0.022 (0.014-0.03)	5.3% (3.5-7.1)	0.021 (0.014-0.029)	5.3% (3.6-7.1)	0.021 (0.015-0.029)	5.4% (3.7-7.3)	0.021 (0.014-0.029)	5.4% (3.6-7.2)
Item-by-Teacher ($var(v_{i(t)})$)	0.012 (0.009-0.015)	2.9% (2.3-3.6)	0.013 (0.01-0.015)	3.1% (2.5-3.8)	0.013 (0.01-0.015)	3.2% (2.5-3.9)	0.013 (0.01-0.015)	3.2% (2.6-3.9)
Item-by-Day ($var(v_{i(d:s:t)})$)	0.067 (0.062-0.07)	16.2% (15.2-17.1)	0.069 (0.065-0.073)	17.2% (16.2-18.2)	0.069 (0.065-0.074)	17.5% (16.5-18.5)	0.07 (0.065-0.074)	17.7% (16.7-18.7)
Item-by-Occasion ($var(v_{i(o:d:s:t)})$)	0.012 (0.007-0.017)	2.9% (1.7-4.1)	0.005 (0.001-0.011)	1.3% (0.3-2.6)	0.005 (0.001-0.011)	1.3% (0.3-2.7)	0 (0-0.005)	0% (0-1.4)
Residual ($var(\epsilon_{ir(o:d:s:t)})$)	0.246 (0.241-0.251)	59.6% (57.7-61.4)	0.245 (0.24-0.25)	61.1% (59.1-62.8)	0.245 (0.24-0.25)	62% (59.9-63.6)	0.25 (0.244-0.253)	63.5% (61.5-64.9)

Note. Each pair of columns shows a separate model. For each regression model, the left column displays the estimated variance and the right column displays the percentage of variance for each error facet.

Appendix B – Fixed Effect Estimates in Scale Score Metric

Table B.1: Fixed Effects for the System Design (SD) Model across the three Instruments in the Scale Score Metric

Names	CLASS	FFT	PLATO
Scored Live (β_{Live})	0.09 (0.06)	0.10 (0.04)**	-0.05 (0.03)
Double Scored (β_{Dbl})	-0.03 (0.04)	0.01 (0.02)	-0.03 (0.02)
Date Scored (m) (β_{DtSc})	-0.02 (0.00)***	-0.01 (0.00)**	-0.01 (0.00)***
Day of the Week (β_{DayWk})			
Tuesday	-0.02 (0.04)	-0.00 (0.02)	-0.02 (0.02)
Wednesday	0.09 (0.04)*	0.05 (0.03)	0.02 (0.02)
Thursday	-0.01 (0.04)	0.03 (0.02)	-0.02 (0.02)
Friday	-0.07 (0.05)	-0.03 (0.03)	-0.00 (0.02)
Observation Month (β_{Month})	-0.03 (0.01)***	-0.02 (0.00)***	-0.01 (0.00)***

Note. Each column shows the results of a separate model for the indicated instrument. Date Scored is scaled so a 1 point difference is one month; Monday is the reference group for the Days of the Week. * p<0.05; ** p<0.01; *** p<0.001.

Table B.2: Fixed Effects for the Curriculum and Instruction (CI) Model across the three Instruments in the Scale Score Metric

Names	CLASS	FFT	PLATO
Scored Live (β_{Live})	0.10 (0.06)	0.09 (0.04)**	-0.04 (0.03)
Double Scored (β_{Dbl})	-0.03 (0.04)	0.01 (0.02)	-0.05 (0.02)**
Date Scored (m) (β_{DtSc})	-0.02 (0.00)***	-0.01 (0.00)**	-0.01 (0.00)***
Day of the Week (β_{DayWk})			
Tuesday	-0.01 (0.04)	0.01 (0.02)	-0.00 (0.02)
Wednesday	0.09 (0.04)*	0.05 (0.03)*	0.01 (0.02)
Thursday	-0.00 (0.04)	0.03 (0.02)	-0.02 (0.02)
Friday	-0.07 (0.05)	-0.03 (0.03)	-0.01 (0.02)
Observation Month (β_{Month})	-0.03 (0.01)***	-0.02 (0.00)***	-0.01 (0.00)*
Content Domain			
Reading (β_{Read})	0.03 (0.05)	-0.04 (0.03)	0.05 (0.02)*
Literature (β_{Lit})	0.10 (0.04)**	0.08 (0.02)***	0.12 (0.02)***
Writing (β_{Write})	0.12 (0.04)***	0.04 (0.02)	0.11 (0.02)***
Grammar ($\beta_{Grammar}$)	0.05 (0.04)	-0.04 (0.02)*	0.00 (0.02)
Interaction Structure			
Discussion (β_{Disc})	0.08 (0.03)**	0.00 (0.02)	0.07 (0.01)***
Independent (β_{Ind})	0.01 (0.05)	0.04 (0.03)	0.05 (0.02)*
Recitation (β_{Rec})	-0.05 (0.03)	-0.00 (0.02)	0.02 (0.02)

Note. Each column shows the results of a separate model for the indicated instrument. Date Scored is scaled so a 1 point difference is one month; Monday is the reference group for the Days of the Week. * p<0.05; ** p<0.01; *** p<0.001.

Table B.3: Fixed Effects for the School Organization (SO) Model across the three Instruments in the Scale Score Metric

Names	CLASS	FFT	PLATO
Scored Live (β_{Live})	0.13 (0.05)*	0.11 (0.03)**	-0.03 (0.03)
Double Scored (β_{Dbl})	-0.04 (0.04)	0.01 (0.02)	-0.05 (0.02)**
Date Scored (m) (β_{DtSc})	-0.01 (0.00)***	-0.00 (0.00)*	-0.01 (0.00)***
Day of the Week (β_{DayWk})			
Tuesday	-0.02 (0.04)	-0.00 (0.02)	-0.01 (0.02)
Wednesday	0.07 (0.04)	0.04 (0.02)	0.00 (0.02)
Thursday	-0.01 (0.04)	0.02 (0.02)	-0.03 (0.02)
Friday	-0.07 (0.05)	-0.04 (0.03)	-0.01 (0.02)
Observation Month (β_{Month})	-0.03 (0.01)***	-0.02 (0.00)***	-0.01 (0.00)**
Content Domain			
Reading (β_{Read})	0.05 (0.05)	-0.03 (0.03)	0.06 (0.02)*
Literature (β_{Lit})	0.07 (0.04)	0.05 (0.02)*	0.10 (0.02)***
Writing (β_{Write})	0.12 (0.03)***	0.03 (0.02)	0.11 (0.02)***
Grammar ($\beta_{Grammar}$)	0.05 (0.04)	-0.04 (0.02)*	0.00 (0.02)
Interaction Structure			
Discussion (β_{Disc})	0.06 (0.03)*	-0.01 (0.02)	0.07 (0.01)***
Independent (β_{Ind})	0.01 (0.04)	0.03 (0.03)	0.05 (0.02)*
Recitation (β_{Rec})	-0.04 (0.03)	-0.00 (0.02)	0.02 (0.02)
Grade			
7th Grade	-0.14 (0.04)**	-0.06 (0.03)*	-0.05 (0.02)*
8th Grade	0.02 (0.04)	0.01 (0.03)	0.01 (0.02)
Prior Ach	0.08 (0.02)**	0.08 (0.01)***	0.02 (0.01)
St. Info Missing	-0.13 (0.07)	-0.07 (0.04)	-0.03 (0.03)
Demo. Composite	-0.09 (0.02)***	-0.03 (0.01)*	-0.03 (0.01)**

Note. Each column shows the results of a separate model for the indicated instrument. Date Scored is scaled so a 1 point difference is one month. Monday is the reference group for the Days of the Week. Sixth grade is the references group for grade. The Demographic Composite represents classrooms that have higher percentages of students who are black, Hispanic, ELL, and FRL. St. Info Missing is a dummy variable indicating if Prior Achievement and Demographic Composite are missing. * p<0.05; ** p<0.01; *** p<0.001.

Table B.4: Item-by-Occasion Fixed Effects for the System Design (SD) Model on the CLASS Instrument in the Scale Score Metric

Item (β_i)	Main	Occasion 2	Occasion 3	Occasion 4+
Positive Climate	4.65 (0.15)***	0.04 (0.03)	0.01 (0.03)	0.00 (0.04)
Negative Climate	6.79 (0.15)***	0.02 (0.03)	0.04 (0.03)	0.06 (0.04)
Adolescent Perspectives	3.05 (0.15)***	0.24 (0.03)***	0.23 (0.03)***	0.28 (0.04)***
Teacher Sensitivity	4.08 (0.15)***	0.10 (0.03)***	0.11 (0.03)***	-0.04 (0.04)
Behavior Management	6.09 (0.15)***	-0.04 (0.03)	-0.05 (0.03)	-0.13 (0.04)**
Productivity	5.79 (0.15)***	0.05 (0.03)	0.07 (0.03)*	0.02 (0.04)
Instructional Learning Formats	3.83 (0.15)***	0.06 (0.03)*	-0.08 (0.03)**	-0.27 (0.04)***
Content Understanding	3.38 (0.15)***	0.07 (0.03)*	-0.10 (0.03)***	-0.29 (0.04)***
Analysis and Problem Solving	2.40 (0.15)***	0.18 (0.03)***	0.17 (0.03)***	0.13 (0.04)**
Quality of Feedback	3.35 (0.15)***	0.18 (0.03)***	0.16 (0.03)***	0.10 (0.04)*
Student Engagement	5.10 (0.15)***	0.04 (0.03)	0.02 (0.03)	0.03 (0.04)

Note. Column 'Main' shows the Item mean on occasion 1; Column '2' shows the deviation of the item on occasion 2; Column '3' shows the deviation of the item on occasion 3; Column '4+' shows the deviation of the item on occasion 4 or higher. * p<0.05; ** p<0.01; *** p<0.001.

Table B.5: Item-by-Occasion Fixed Effects for the System Design (SD) Model on the PLATO Instrument in the Scale Score Metric

Item (β_i)	Main	Occasion 2	Occasion 3	Occasion 4+
Purpose	2.93 (0.07)***	0.01 (0.02)	-0.00 (0.02)	-0.03 (0.03)
Intellectual Challenge	2.09 (0.07)***	0.12 (0.02)***	0.14 (0.02)***	0.12 (0.03)***
Representation of Content	2.45 (0.07)***	0.08 (0.02)***	0.02 (0.02)	-0.10 (0.03)**
Connections to Prior Knowledge	1.73 (0.07)***	-0.16 (0.02)***	-0.32 (0.02)***	-0.47 (0.03)***
Connections to Personal Experience	1.36 (0.07)***	0.04 (0.02)*	-0.01 (0.02)	-0.05 (0.03)
Explicit Strategy Instruction	1.24 (0.07)***	0.02 (0.02)	-0.03 (0.02)	-0.05 (0.03)
Modeling	1.25 (0.07)***	0.10 (0.02)***	0.06 (0.02)*	0.01 (0.03)
Guided Practice	2.35 (0.07)***	0.13 (0.02)***	0.25 (0.02)***	0.24 (0.03)***
Classroom Discourse	2.05 (0.07)***	0.14 (0.02)***	0.11 (0.02)***	0.08 (0.03)*
Text Based Instruction	1.78 (0.07)***	0.29 (0.02)***	0.33 (0.02)***	0.33 (0.03)***
Acc. for Language Learning	1.44 (0.07)***	0.02 (0.02)	-0.05 (0.02)*	-0.11 (0.03)***
Behavior Management	3.98 (0.07)***	-0.03 (0.02)	-0.03 (0.02)	-0.02 (0.03)
Time Management	3.75 (0.07)***	0.09 (0.02)***	0.12 (0.02)***	0.14 (0.03)***

Note. Column 'Main' shows the Item mean on occasion 1; Column '2' shows the deviation of the item on occasion 2; Column '3' shows the deviation of the item on occasion 3; Column '4+' shows the deviation of the item on occasion 4 or higher. Acc. for Language Learn is Accommodations for Language Learning. * p<0.05; ** p<0.01; *** p<0.001.

Appendix C – Numeric Results of Score Reliabilities across Models

Table C.1: Estimated Reliability for Each Model for the Listed Number Days Scores and Raters Scoring each Day

Instrument	Raters	Days	Base	SD	CI	SO
CLASS	1	1	0.23 (0.17-0.30)	0.24 (0.18-0.30)	0.22 (0.17-0.29)	0.14 (0.09-0.19)
CLASS	1	2	0.38 (0.29-0.46)	0.38 (0.30-0.47)	0.36 (0.29-0.44)	0.24 (0.16-0.32)
CLASS	1	3	0.48 (0.38-0.56)	0.48 (0.39-0.57)	0.46 (0.37-0.55)	0.32 (0.22-0.41)
CLASS	1	4	0.55 (0.45-0.63)	0.55 (0.47-0.64)	0.53 (0.44-0.62)	0.39 (0.28-0.48)
CLASS	1	6	0.64 (0.55-0.72)	0.65 (0.57-0.72)	0.63 (0.54-0.71)	0.48 (0.36-0.58)
CLASS	1	8	0.71 (0.62-0.77)	0.71 (0.64-0.78)	0.69 (0.61-0.76)	0.55 (0.43-0.65)
CLASS	1.2	1	0.26 (0.19-0.33)	0.27 (0.20-0.33)	0.25 (0.19-0.32)	0.16 (0.10-0.21)
CLASS	1.2	2	0.41 (0.32-0.50)	0.42 (0.34-0.50)	0.40 (0.32-0.48)	0.27 (0.18-0.35)
CLASS	1.2	3	0.51 (0.42-0.60)	0.52 (0.43-0.60)	0.50 (0.41-0.58)	0.35 (0.25-0.45)
CLASS	1.2	4	0.58 (0.49-0.66)	0.59 (0.50-0.67)	0.57 (0.48-0.65)	0.42 (0.30-0.52)
CLASS	1.2	6	0.68 (0.59-0.75)	0.68 (0.60-0.75)	0.66 (0.58-0.74)	0.52 (0.40-0.62)
CLASS	1.2	8	0.74 (0.66-0.80)	0.74 (0.67-0.80)	0.72 (0.65-0.79)	0.59 (0.47-0.68)
CLASS	2	1	0.34 (0.26-0.42)	0.35 (0.27-0.42)	0.33 (0.25-0.40)	0.21 (0.14-0.29)
CLASS	2	2	0.51 (0.41-0.59)	0.51 (0.42-0.60)	0.49 (0.40-0.57)	0.35 (0.24-0.45)
CLASS	2	3	0.61 (0.51-0.69)	0.61 (0.52-0.69)	0.59 (0.50-0.67)	0.44 (0.33-0.55)
CLASS	2	4	0.67 (0.58-0.74)	0.68 (0.59-0.75)	0.66 (0.57-0.73)	0.52 (0.39-0.62)
CLASS	2	6	0.76 (0.68-0.81)	0.76 (0.69-0.82)	0.74 (0.67-0.80)	0.61 (0.49-0.71)
CLASS	2	8	0.80 (0.74-0.85)	0.81 (0.74-0.85)	0.79 (0.73-0.84)	0.68 (0.56-0.76)
FFT	1	1	0.26 (0.20-0.33)	0.25 (0.19-0.31)	0.23 (0.18-0.30)	0.14 (0.09-0.19)
FFT	1	2	0.42 (0.33-0.49)	0.40 (0.32-0.48)	0.38 (0.30-0.46)	0.24 (0.16-0.32)
FFT	1	3	0.52 (0.43-0.60)	0.50 (0.42-0.58)	0.48 (0.39-0.56)	0.32 (0.23-0.41)
FFT	1	4	0.59 (0.50-0.66)	0.57 (0.49-0.65)	0.55 (0.46-0.63)	0.39 (0.28-0.48)
FFT	1	6	0.68 (0.60-0.75)	0.67 (0.59-0.73)	0.64 (0.56-0.72)	0.49 (0.37-0.58)
FFT	1	8	0.74 (0.66-0.80)	0.73 (0.66-0.79)	0.71 (0.63-0.77)	0.56 (0.44-0.65)
FFT	1.2	1	0.30 (0.22-0.37)	0.28 (0.22-0.35)	0.26 (0.20-0.33)	0.16 (0.10-0.22)
FFT	1.2	2	0.45 (0.37-0.54)	0.44 (0.36-0.52)	0.42 (0.33-0.50)	0.27 (0.19-0.36)
FFT	1.2	3	0.55 (0.46-0.63)	0.54 (0.46-0.62)	0.52 (0.43-0.60)	0.36 (0.26-0.46)
FFT	1.2	4	0.62 (0.54-0.70)	0.61 (0.53-0.68)	0.59 (0.50-0.66)	0.43 (0.32-0.53)
FFT	1.2	6	0.71 (0.63-0.78)	0.70 (0.63-0.76)	0.68 (0.60-0.75)	0.53 (0.41-0.63)
FFT	1.2	8	0.77 (0.70-0.82)	0.76 (0.69-0.81)	0.74 (0.67-0.80)	0.60 (0.48-0.69)
FFT	2	1	0.39 (0.30-0.47)	0.38 (0.30-0.45)	0.35 (0.28-0.43)	0.23 (0.15-0.30)
FFT	2	2	0.56 (0.47-0.64)	0.54 (0.46-0.62)	0.52 (0.43-0.60)	0.37 (0.26-0.46)
FFT	2	3	0.65 (0.57-0.73)	0.64 (0.56-0.71)	0.62 (0.53-0.70)	0.46 (0.34-0.57)
FFT	2	4	0.72 (0.64-0.78)	0.70 (0.63-0.77)	0.68 (0.60-0.75)	0.53 (0.41-0.63)
FFT	2	6	0.79 (0.72-0.84)	0.78 (0.72-0.83)	0.76 (0.70-0.82)	0.63 (0.51-0.72)
FFT	2	8	0.83 (0.78-0.88)	0.83 (0.77-0.87)	0.81 (0.75-0.86)	0.69 (0.58-0.78)
PLATO	1	1	0.22 (0.15-0.28)	0.20 (0.14-0.26)	0.17 (0.11-0.23)	0.13 (0.08-0.18)
PLATO	1	2	0.35 (0.26-0.44)	0.33 (0.25-0.41)	0.29 (0.21-0.37)	0.23 (0.14-0.31)
PLATO	1	3	0.45 (0.35-0.54)	0.43 (0.33-0.51)	0.38 (0.28-0.47)	0.31 (0.20-0.40)
PLATO	1	4	0.52 (0.42-0.61)	0.50 (0.40-0.59)	0.45 (0.34-0.54)	0.37 (0.25-0.47)
PLATO	1	6	0.62 (0.52-0.70)	0.60 (0.50-0.68)	0.55 (0.44-0.64)	0.47 (0.34-0.58)

PLATO	1	8	0.68 (0.59-0.76)	0.66 (0.57-0.74)	0.61 (0.51-0.70)	0.54 (0.40-0.64)
PLATO	1.2	1	0.24 (0.17-0.30)	0.22 (0.16-0.28)	0.19 (0.13-0.25)	0.14 (0.09-0.20)
PLATO	1.2	2	0.38 (0.29-0.47)	0.36 (0.27-0.44)	0.31 (0.23-0.40)	0.25 (0.16-0.34)
PLATO	1.2	3	0.48 (0.38-0.57)	0.46 (0.36-0.54)	0.41 (0.30-0.50)	0.33 (0.22-0.43)
PLATO	1.2	4	0.55 (0.45-0.64)	0.53 (0.43-0.61)	0.48 (0.37-0.57)	0.40 (0.28-0.50)
PLATO	1.2	6	0.65 (0.55-0.72)	0.63 (0.53-0.70)	0.58 (0.47-0.67)	0.50 (0.36-0.60)
PLATO	1.2	8	0.71 (0.62-0.78)	0.69 (0.60-0.76)	0.64 (0.54-0.73)	0.57 (0.43-0.67)
PLATO	2	1	0.29 (0.22-0.37)	0.27 (0.20-0.35)	0.24 (0.16-0.31)	0.19 (0.11-0.26)
PLATO	2	2	0.45 (0.35-0.54)	0.43 (0.33-0.52)	0.38 (0.28-0.48)	0.31 (0.21-0.41)
PLATO	2	3	0.55 (0.45-0.64)	0.53 (0.42-0.62)	0.48 (0.37-0.58)	0.40 (0.28-0.51)
PLATO	2	4	0.62 (0.52-0.70)	0.60 (0.49-0.68)	0.55 (0.44-0.64)	0.47 (0.34-0.58)
PLATO	2	6	0.71 (0.62-0.78)	0.69 (0.59-0.76)	0.65 (0.54-0.73)	0.57 (0.44-0.67)
PLATO	2	8	0.77 (0.69-0.83)	0.75 (0.66-0.81)	0.71 (0.61-0.78)	0.64 (0.51-0.73)

Appendix D – Item Specific Variance Components

Table D.1: Item-Specific Variance Components from Base GTheory Model for Instrument CLASS

Facet	PC	NC	RSP	TS	BM	PD	ILF	CU	APS	QF	ENG
Teacher ($var(v_t)$)	0.159 13.1%	0.024 8.4%	0.184 12%	0.165 11.6%	0.052 9.3%	0.042 5.8%	0.154 11.8%	0.107 8%	0.055 5%	0.109 7.2%	0.08 9%
Day ($var(v_{d:s:t})$)	0.056 4.7%	0.012 4.3%	0.108 7%	0.064 4.5%	0.068 12.2%	0.055 7.5%	0.017 1.3%	0.114 8.6%	0.054 4.9%	0.068 4.4%	0.038 4.3%
Occasion ($var(v_{o:d:s:t})$)	0.014 1.2%	0.019 6.6%	0.118 7.7%	0.038 2.7%	0.017 3%	0.064 8.8%	0.096 7.4%	0.153 11.4%	0.046 4.1%	0.147 9.6%	0.055 6.1%
Rater ($var(v_r)$)	0.395 32.7%	0.058 20%	0.319 20.7%	0.432 30.4%	0.08 14.3%	0.143 19.5%	0.261 20.1%	0.234 17.5%	0.35 31.4%	0.385 25.2%	0.24 26.8%
Rater-by-Teacher ($var(v_{rt})$)	0.067 5.6%	0.017 5.9%	0.007 0.4%	0.014 1%	0 0%	0 0%	0 0%	0.006 0.4%	0 0%	0 0%	0.004 0.5%
Rater-by-Day ($var(v_{r(d:s:t)})$)	0.266 22%	0.074 25.5%	0.389 25.2%	0.307 21.6%	0.143 25.6%	0.195 26.7%	0.404 31%	0.317 23.7%	0.293 26.3%	0.388 25.4%	0.245 27.4%
Residual ($var(\epsilon_{ir(o:d:s:t)})$)	0.25 20.7%	0.085 29.4%	0.416 27%	0.403 28.3%	0.198 35.6%	0.232 31.7%	0.369 28.4%	0.405 30.3%	0.316 28.3%	0.431 28.2%	0.233 26%

Note. Separate regressions were run for each item. For each regression model, the estimated variance is shown above the percentage of variance for each error facet. PC=Positive Climate; NC=Negative Climate; RSP=Regard for Adolescent Behavior; TS=Teacher Sensitivity; BM=Behavior Management; PD=Productivity; ILF=Instructional Learning Formats; CU=Content Understanding; APS=Analysis and Problem Solving; QF=Quality of Feedback; ENG=Student Engagement. Negative Climate has been reverse coded so higher scores capture higher quality.

Table D.2: Item-Specific Variance Components from Base GTheory Model for Instrument FFT

Facet	RR	CL	MCP	MSB	OPS	CS	KC	QDT	ESL	UAI	FR
Teacher ($var(v_t)$)	0.032	0.064	0.05	0.045	0.029	0.034	0.038	0.038	0.044	0.011	0.019
	17.6%	19.2%	15.8%	20.8%	11.5%	12.5%	12.3%	14.8%	12.7%	4.2%	7.4%
Day ($var(v_{d:s:t})$)	0.013	0.019	0.023	0.022	0.028	0.024	0.011	0.017	0.033	0.02	0.006
	6.9%	5.7%	7.4%	10.2%	11.2%	8.9%	3.6%	6.8%	9.5%	7.7%	2.3%
Rater ($var(v_r)$)	0.005	0.027	0.022	0.01	0.006	0.021	0.036	0.014	0.053	0.035	0.022
	2.6%	8%	6.9%	4.4%	2.5%	7.7%	11.6%	5.4%	15.3%	13.4%	8.4%
Rater-by-Teacher	0	0.008	0.009	0	0.018	0.002	0.018	0.002	0.025	0.007	0.014
($var(v_{rt})$)	0%	2.3%	2.8%	0%	7.3%	0.7%	5.9%	0.6%	7.2%	2.7%	5.3%
Rater-by-Day	0.069	0.093	0.041	0.053	0.074	0.072	0.109	0.05	0.067	0.077	0.093
($var(v_{r(d:s:t)})$)	37.9%	27.8%	13%	24.7%	29.5%	26.5%	35%	19.6%	19.1%	29.1%	36.2%
Residual	0.064	0.123	0.172	0.085	0.096	0.118	0.099	0.134	0.126	0.113	0.104
($var(\epsilon_{ir(o:d:s:t)})$)	35.1%	37%	54.1%	39.8%	38%	43.6%	31.6%	52.7%	36.1%	42.9%	40.4%

Note. Separate regressions were run for each item. For each regression model, the estimated variance is shown above the percentage of variance for each error facet. RR=Respect and Rapport; CL=Culture for Learning; MCP=Managing Classroom Procedures; MSB=Managing Student Behavior; OPS=Organizing Physical Space; CS=Communicating with Students; KC=Knowledge of Content and Pedagogy; QDT=Questioning Discussion Techniques; ESL=Engaging Students in Learning; UAI=Using Assessment in Instruction; FR=Flexibility and Responsiveness.

Table D.3: Item-Specific Variance Components from Base GTheory Model for Instrument PLATO

Facet	PURP	INTC	RC	CPK	CPE	ESI	MOD	GP	CD	TBI	ALL	BMN	TMN
Teacher $(var(v_t))$	0.009	0.021	0.038	0.022	0.015	0.006	0.003	0.018	0.044	0.086	0.01	0.008	0.014
	3.9%	5.6%	8.4%	4.2%	4%	3.1%	0.9%	2.4%	11.2%	8%	3.6%	7.2%	5.4%
Day $(var(v_{d:s:t}))$	0.018	0.028	0.034	0	0.054	0.008	0.013	0.046	0.035	0.198	0.006	0.016	0.004
	7.4%	7.8%	7.4%	0%	14.1%	4%	3.9%	6%	8.8%	18.4%	2.1%	13.9%	1.8%
Occasion	0.027	0.031	0.082	0.089	0.074	0.018	0.064	0.108	0.044	0.154	0.017	0.01	0.034
$(var(v_{o:d:s:t}))$	11.3%	8.6%	18.1%	17.1%	19.4%	9.6%	19.5%	14.3%	11.2%	14.3%	5.7%	8.5%	13.6%
Rater $(var(v_r))$	0.018	0.048	0.012	0.024	0.014	0.018	0.018	0.073	0.03	0.021	0.035	0.001	0.006
	7.3%	13.1%	2.7%	4.6%	3.7%	9.5%	5.6%	9.7%	7.7%	2%	11.9%	0.4%	2.4%
Rater-by-Teacher	0.008	0.014	0.007	0.003	0.022	0	0.018	0	0.002	0.031	0	0	0
$(var(v_{rt}))$	3.5%	3.9%	1.4%	0.5%	5.8%	0%	5.6%	0%	0.6%	2.9%	0%	0%	0%
Rater-by-Day	0.04	0.053	0.058	0.065	0.004	0.045	0.04	0.141	0.077	0.191	0.086	0.019	0.055
$(var(v_{r(d:s:t)}))$	16.8%	14.4%	12.7%	12.5%	1%	23.4%	12.2%	18.6%	19.7%	17.8%	29.6%	16.4%	22%
Residual	0.119	0.171	0.224	0.319	0.199	0.097	0.17	0.371	0.161	0.392	0.137	0.063	0.137
$(var(\epsilon_{ir(o:d:s:t)}))$	49.9%	46.7%	49.3%	61.1%	51.9%	50.4%	52.3%	49%	40.9%	36.5%	47.2%	53.6%	54.9%

Note. Separate regressions were run for each item. For each regression model, the estimated variance is shown above the percentage of variance for each error facet. PURP=Purpose; INTC=Intellectual Climate; RC=Representation of Content; CPK=Connections to Prior Knowledge; CPE=Connections to Personal and /or Cultural Experience; ESI=Explicit Strategy Instruction; MOD=Modeling; GP=Guided Practice; CD=Classroom Discussion; TBI=Text-Based Instruction; ALL=Accommodations for Language Learners; BMN=Behavior Management; TMN=Time Management.

Table D.4: Item-Specific Variance Components from SO GTheory Model for Instrument CLASS

Facet	PC	NC	RSP	TS	BM	PD	ILF	CU	APS	QF	ENG
Teacher ($var(v_t)$)	0.074 6.8%	0.02 7%	0.055 4.3%	0.072 5.6%	0.044 8.3%	0.023 3.2%	0.064 5.9%	0.052 4.5%	0.023 2.5%	0.046 3.4%	0.037 4.5%
Day ($var(v_{d:s:t})$)	0.053 4.8%	0.007 2.5%	0.113 8.8%	0.055 4.3%	0.057 10.7%	0.049 6.9%	0 0%	0.092 8%	0.041 4.4%	0.036 2.6%	0.037 4.5%
Occasion ($var(v_{o:d:s:t})$)	0.015 1.3%	0.019 6.9%	0.102 7.9%	0.035 2.8%	0.015 2.8%	0.065 9.1%	0.087 8%	0.136 11.8%	0.037 4%	0.14 10.2%	0.055 6.7%
Rater ($var(v_r)$)	0.398 36.3%	0.058 20.7%	0.269 20.9%	0.42 33%	0.081 15.2%	0.148 20.8%	0.203 18.6%	0.178 15.4%	0.264 28.2%	0.349 25.5%	0.223 27.2%
Rater-by-Teacher ($var(v_{rt})$)	0.088 8%	0.021 7.5%	0.057 4.4%	0.059 4.6%	0 0%	0 0%	0 0%	0.04 3.5%	0 0%	0.013 1%	0.009 1.1%
Rater-by-Day ($var(v_{r(d:s:t)})$)	0.219 20%	0.071 25.3%	0.281 21.8%	0.229 18%	0.139 26.2%	0.194 27.4%	0.367 33.7%	0.252 21.8%	0.257 27.4%	0.351 25.7%	0.227 27.6%
Residual ($var(\epsilon_{ir(o:d:s:t)})$)	0.249 22.7%	0.084 30%	0.41 31.9%	0.402 31.6%	0.196 36.8%	0.231 32.6%	0.369 33.9%	0.406 35.1%	0.314 33.6%	0.431 31.5%	0.233 28.3%

Note. Separate regressions were run for each item. For each regression model, the estimated variance is shown above the percentage of variance for each error facet. PC=Positive Climate; NC=Negative Climate; RSP=Regard for Adolescent Behavior; TS=Teacher Sensitivity; BM=Behavior Management; PD=Productivity; ILF=Instructional Learning Formats; CU=Content Understanding; APS=Analysis and Problem Solving; QF=Quality of Feedback; ENG=Student Engagement. Negative Climate has been reverse coded so higher scores capture higher quality.

Table D.5: Item-Specific Variance Components from SO GTheory Model for Instrument FFT

Facet	RR	CL	MCP	MSB	OPS	CS	KC	QDT	ESL	UAI	FR
Teacher ($var(v_t)$)	0.025 14.8%	0.026 9.5%	0.032 10.8%	0.036 17.9%	0.01 4.7%	0.014 5.5%	0.008 3.2%	0.017 7.5%	0.015 4.9%	0.003 1.2%	0.003 1.3%
Day ($var(v_{d:s:t})$)	0.011 6.3%	0.016 5.8%	0.025 8.3%	0.02 9.7%	0.029 13.2%	0.026 10.3%	0.014 5.3%	0.015 6.8%	0.027 8.8%	0.017 6.8%	0.008 3.5%
Rater ($var(v_r)$)	0.005 2.9%	0.019 7.1%	0.022 7.4%	0.009 4.6%	0.004 1.9%	0.022 8.8%	0.029 11.6%	0.011 5.1%	0.048 15.8%	0.036 14.4%	0.022 9.6%
Rater-by-Teacher ($var(v_{rt})$)	0 0%	0.014 5.1%	0.015 5%	0 0%	0.019 8.4%	0.001 0.4%	0.016 6.4%	0 0%	0.025 8.1%	0 0.2%	0.015 6.8%
Rater-by-Day ($var(v_{r(d:s:t)})$)	0.066 38.3%	0.075 27.7%	0.033 11%	0.056 27.3%	0.061 27.8%	0.077 30.9%	0.097 38.1%	0.047 21.1%	0.069 22.7%	0.081 32.6%	0.072 32.3%
Residual ($var(\epsilon_{ir(o:d:s:t)})$)	0.065 37.7%	0.122 44.8%	0.172 57.5%	0.083 40.6%	0.097 44.1%	0.109 44.1%	0.09 35.4%	0.133 59.6%	0.121 39.7%	0.112 44.8%	0.104 46.5%

Note. Separate regressions were run for each item. For each regression model, the estimated variance is shown above the percentage of variance for each error facet. RR=Respect and Rapport; CL=Culture for Learning; MCP=Managing Classroom Procedures; MSB=Managing Student Behavior; OPS=Organizing Physical Space; CS=Communicating with Students; KC=Knowledge of Content and Pedagogy; QDT=Questioning Discussion Techniques; ESL=Engaging Students in Learning; UAI=Using Assessment in Instruction; FR=Flexibility and Responsiveness.

Table D.6: Item-Specific Variance Components from SO GTheory Model for Instrument PLATO

Facet	PURP	INTC	RC	CPK	CPE	ESI	MOD	GP	CD	TBI	ALL	BMN	TMN
Teacher ($var(\upsilon_t)$)	0.004 1.9%	0.006 1.9%	0.022 5.1%	0.016 3.4%	0.007 1.8%	0.006 3.3%	0.002 0.6%	0.015 2%	0.012 3.7%	0.046 5.3%	0.003 1.2%	0.008 6.5%	0.01 4.1%
Day ($var(\upsilon_{d:s:t})$)	0.014 6.2%	0.024 7.1%	0.028 6.7%	0.002 0.4%	0.051 13.9%	0.011 5.9%	0.014 4.7%	0.037 5%	0.028 8.2%	0.069 7.8%	0.008 3.2%	0.015 12.9%	0.006 2.6%
Occasion ($var(\upsilon_{o:d:s:t})$)	0.027 11.6%	0.025 7.3%	0.079 18.6%	0.067 13.9%	0.074 20.1%	0.018 9.9%	0.061 20.1%	0.093 12.5%	0.038 11.3%	0.117 13.3%	0.014 5.5%	0.01 8.4%	0.03 12.4%
Rater ($var(\upsilon_r)$)	0.017 7.4%	0.046 13.7%	0.008 2%	0.015 3%	0.012 3.4%	0.008 4.3%	0.007 2.3%	0.09 12.1%	0.028 8.5%	0.023 2.6%	0.018 7.1%	0 0.4%	0.008 3.2%
Rater-by-Teacher ($var(\upsilon_{rt})$)	0.009 3.9%	0.02 5.9%	0.009 2%	0.001 0.2%	0.02 5.5%	0 0%	0.024 7.8%	0 0%	0.007 2%	0.005 0.6%	0 0%	0 0%	0 0%
Rater-by-Day ($var(\upsilon_{r(d:s:t)})$)	0.038 16.6%	0.043 12.7%	0.054 12.8%	0.071 14.7%	0.005 1.2%	0.038 21.3%	0.025 8.2%	0.135 18.1%	0.059 17.8%	0.223 25.3%	0.073 28.6%	0.02 17%	0.049 20.3%
Residual ($var(\epsilon_{ir(o:d:s:t)})$)	0.12 52.4%	0.173 51.3%	0.224 52.9%	0.311 64.4%	0.199 54.1%	0.097 55.2%	0.17 56.3%	0.373 50.3%	0.162 48.5%	0.398 45.2%	0.138 54.4%	0.063 54.8%	0.138 57.3%

Note. Separate regressions were run for each item. For each regression model, the estimated variance is shown above the percentage of variance for each error facet. PURP=Purpose; INTC=Intellectual Climate; RC=Representation of Content; CPK=Connections to Prior Knowledge; CPE=Connections to Personal and /or Cultural Experience; ESI=Explicit Strategy Instruction; MOD=Modeling; GP=Guided Practice; CD=Classroom Discussion; TBI=Text-Based Instruction; ALL=Accommodations for Language Learners; BMN=Behavior Management; TMN=Time Management.

Table D.7: Item-Specific Fixed Effects from SD GTheory Model for Instrument CLASS in Scale Score Metric

Facet	PC	NC	RSP	TS	BM	PD	ILF	CU	APS	QF	ENG
Intercept	4.64***	6.80***	2.96***	4.02***	6.16***	5.85***	3.76***	3.36***	2.46***	3.35***	5.05***
	(0.20)	(0.08)	(0.17)	(0.20)	(0.10)	(0.13)	(0.15)	(0.15)	(0.16)	(0.19)	(0.15)
Scored Live	0.22*	0.06	0.06	-0.14	-0.03	0.14	0.08	0.16	0.32***	-0.15	0.32***
(β_{Live})	(0.09)	(0.05)	(0.10)	(0.10)	(0.07)	(0.08)	(0.10)	(0.10)	(0.09)	(0.10)	(0.08)
Double Scored	-0.13*	-0.01	-0.02	-0.08	0.02	0.01	-0.01	-0.05	-0.04	-0.11	0.05
(β_{Dbl})	(0.06)	(0.03)	(0.07)	(0.06)	(0.04)	(0.05)	(0.06)	(0.06)	(0.06)	(0.07)	(0.05)
Date Scored (m)	-0.01	0.01**	-0.05***	-0.04***	0.01*	0.02***	-0.04***	-0.03***	-0.02**	-0.04***	0.00
(β_{DtSc})	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Day of the Week (β_{DayWk})											
Tuesday	0.00	0.03	-0.02	-0.01	0.09	0.01	-0.01	-0.05	-0.05	-0.13	0.03
	(0.06)	(0.03)	(0.07)	(0.07)	(0.05)	(0.05)	(0.07)	(0.07)	(0.06)	(0.07)	(0.06)
Wednesday	0.11	0.05	0.17*	0.07	0.11*	0.06	0.12	0.08	0.07	0.12	0.18**
	(0.07)	(0.04)	(0.08)	(0.07)	(0.05)	(0.06)	(0.07)	(0.08)	(0.06)	(0.08)	(0.06)
Thursday	0.08	0.02	0.06	0.05	0.03	0.03	0.04	-0.13	-0.05	-0.04	0.04
	(0.06)	(0.03)	(0.08)	(0.07)	(0.05)	(0.06)	(0.07)	(0.07)	(0.06)	(0.07)	(0.06)
Friday	-0.07	0.00	-0.01	-0.10	0.03	0.01	-0.08	-0.18*	-0.11	-0.18*	0.05
	(0.08)	(0.04)	(0.09)	(0.08)	(0.06)	(0.07)	(0.08)	(0.09)	(0.07)	(0.09)	(0.07)
Observation	-0.04***	-0.02***	-0.04**	-0.03*	-0.04***	-0.03***	-0.04***	-0.03*	-0.05***	-0.03*	-0.03***
Month (β_{Month})	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Segment											
2	0.04	0.02	0.23***	0.10***	-0.04*	0.05*	0.06*	0.06*	0.17***	0.18***	0.04
	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)
3	0.02	0.01	0.27***	0.12***	-0.10***	0.03	-0.06	-0.08*	0.20***	0.16***	0.02
	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)
4+	0.03	0.01	0.34***	-0.02	-0.19***	-0.04	-0.23***	-0.28***	0.16***	0.11*	0.03
	(0.03)	(0.02)	(0.05)	(0.04)	(0.03)	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.03)

Note. Each column shows the results of a separate model for the indicated item and instrument. Date Scored is scaled so a 1 point difference is one month; Monday is the reference group for the Days of the Week. PC=Positive Climate; NC=Negative Climate; RSP=Regard for Adolescent Behavior; TS=Teacher Sensitivity; BM=Behavior Management; PD=Productivity; ILF=Instructional Learning Formats; CU=Content Understanding; APS=Analysis and Problem Solving; QF=Quality of Feedback; ENG=Student Engagement. Negative Climate has been reverse coded so higher scores capture higher quality.* p<0.05; ** p<0.01; *** p<0.001.

Table D.8: Item-Specific Fixed Effects from SD GTheory Model for Instrument FFT in Scale Score Metric

Facet	RR	CL	MCP	MSB	OPS	CS	KC	QDT	ESL	UAI	FR
Intercept	2.90***	2.40***	2.56***	2.86***	2.34***	2.69***	2.24***	2.04***	2.32***	2.00***	2.19***
	(0.04)	(0.06)	(0.07)	(0.05)	(0.05)	(0.06)	(0.07)	(0.06)	(0.08)	(0.07)	(0.06)
Scored Live	0.09	0.30***	-0.02	-0.06	-0.12	-0.04	0.20**	0.18**	0.14*	-0.03	0.24***
(β_{Live})	(0.05)	(0.07)	(0.07)	(0.05)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)
Double Scored	-0.01	-0.01	-0.06	0.00	0.00	0.02	0.03	0.04	0.02	0.07	0.03
(β_{Dbl})	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Date Scored (m)	0.01	-0.00	-0.01	-0.00	-0.02***	-0.01***	-0.01**	-0.00	-0.00	-0.01***	-0.01*
(β_{DtSc})	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Day of the Week (β_{DayWk})											
Tuesday	-0.01	0.02	0.03	0.01	0.01	0.04	-0.02	-0.02	0.00	-0.01	-0.01
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Wednesday	0.07*	0.11*	0.07	0.05	0.06	0.07	0.06	0.04	0.08	0.06	0.05
	(0.04)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)
Thursday	0.04	0.09*	0.02	0.02	0.09*	0.00	0.02	-0.01	0.05	0.04	0.02
	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)
Friday	0.01	-0.04	-0.01	-0.01	-0.01	-0.03	-0.05	-0.02	-0.03	-0.06	0.01
	(0.04)	(0.05)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)
Observation	-0.02***	-0.03***	-0.02*	-0.02**	-0.01	-0.02**	-0.01	-0.02**	-0.02**	-0.00	-0.02***
Month (β_{Month})	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Segment 2	-0.01	-0.03	-0.03	-0.07*	0.01	-0.09*	-0.10**	-0.03	-0.07	0.00	-0.00
	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)

Note. Each column shows the results of a separate model for the indicated item and instrument. Date Scored is scaled so a 1 point difference is one month; Monday is the reference group for the Days of the Week. RR=Respect and Rapport; CL=Culture for Learning; MCP=Managing Classroom Procedures; MSB=Managing Student Behavior; OPS=Organizing Physical Space; CS=Communicating with Students; KC=Knowledge of Content and Pedagogy; QDT=Questioning Discussion Techniques; ESL=Engaging Students in Learning; UAI=Using Assessment in Instruction; FR=Flexibility and Responsiveness. * p<0.05; ** p<0.01; *** p<0.001.

Table D.9: Item-Specific Fixed Effects from SD GTheory Model for Instrument PLATO in Scale Score Metric

Facet	PURP	INTC	RC	CPK	CPE	ESI	MOD	GP	CD	TBI	ALL	BMN	TMN
Intercept	2.88***	2.14***	2.50***	1.69***	1.29***	1.17***	1.27***	2.47***	2.06***	1.76***	1.36***	3.97***	3.81***
	(0.06)	(0.10)	(0.07)	(0.07)	(0.06)	(0.05)	(0.05)	(0.14)	(0.08)	(0.10)	(0.07)	(0.03)	(0.05)
Scored Live	0.10*	-0.04	-0.07	-0.10	-0.17***	0.10*	0.14**	-0.22**	-0.26***	-0.17	0.08	0.02	-0.02
(β_{Live})	(0.05)	(0.05)	(0.06)	(0.06)	(0.05)	(0.04)	(0.05)	(0.08)	(0.06)	(0.10)	(0.05)	(0.03)	(0.05)
Double	-0.04	-0.04	-0.06	0.01	0.01	-0.02	-0.06*	-0.05	-0.01	-0.04	-0.03	0.01	0.03
Scored (β_{Dbl})	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.02)	(0.03)
Date Scored	-0.00	-0.01**	-0.00	-0.02***	-0.02***	-0.01*	-0.00	-0.00	-0.03***	-0.01*	-0.02***	0.00	0.01*
(m) (β_{DtSc})	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Day of the Week (β_{DayWk})													
Tuesday	0.03	-0.01	-0.01	-0.00	0.03	-0.01	-0.06	0.03	-0.05	-0.10	-0.05	0.04	0.02
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.05)	(0.04)	(0.07)	(0.03)	(0.02)	(0.03)
Wednesday	0.07*	0.06	0.02	0.00	0.06	-0.03	-0.08*	0.08	0.04	0.13	-0.05	0.03	0.03
	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)	(0.03)	(0.04)	(0.06)	(0.04)	(0.08)	(0.03)	(0.02)	(0.03)
Thursday	-0.03	-0.00	-0.03	-0.02	0.04	-0.03	-0.06	-0.04	-0.02	0.04	-0.05	0.00	-0.01
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.05)	(0.04)	(0.08)	(0.03)	(0.02)	(0.03)
Friday	0.02	0.02	0.03	0.01	0.06	-0.06	-0.07	0.01	-0.05	0.02	-0.03	-0.02	0.02
	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.03)	(0.04)	(0.07)	(0.05)	(0.09)	(0.04)	(0.03)	(0.04)
Obs. Month	-0.01	-0.03***	-0.01*	-0.01	-0.01	-0.00	-0.01**	-0.02**	-0.02*	-0.01	-0.00	-0.01**	-0.02***
(β_{Month})	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Segment													
2	0.01	0.12***	0.08***	-0.16***	0.05*	0.02	0.10***	0.13***	0.14***	0.29***	0.02	-0.03**	0.10***
	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)
3	0.00	0.15***	0.01	-0.31***	-0.01	-0.02	0.08***	0.22***	0.12***	0.33***	-0.02	-0.04**	0.10***
	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)
4+	-0.03	0.12***	-0.11**	-0.45***	-0.05	-0.04	0.03	0.20***	0.10***	0.34***	-0.08**	-0.04*	0.10***
	(0.02)	(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.04)	(0.03)	(0.05)	(0.03)	(0.02)	(0.03)

Note. Each column shows the results of a separate model for the indicated item and instrument. Date Scored is scaled so a 1 point difference is one month; Monday is the reference group for the Days of the Week. PURP=Purpose; INTC=Intellectual Climate; RC=Representation of Content; CPK=Connections to Prior Knowledge; CPE=Connections to Personal and /or Cultural Experience; ESI=Explicit Strategy Instruction; MOD=Modeling; GP=Guided Practice; CD=Classroom Discussion; TBI=Text-Based Instruction; ALL=Accommodations for Language Learners; BMN=Behavior Management; TMN=Time Management. * p<0.05; ** p<0.01; *** p<0.001.

Table D.10: Item-Specific Fixed Effects from CI GTheory Model for Instrument CLASS in Scale Score Metric

Facet	PC	NC	RSP	TS	BM	PD	ILF	CU	APS	QF	ENG
Intercept	4.53***	6.78***	2.85***	3.93***	6.14***	5.82***	3.60***	3.12***	2.34***	3.09***	5.02***
	(0.20)	(0.08)	(0.18)	(0.21)	(0.11)	(0.14)	(0.16)	(0.16)	(0.17)	(0.20)	(0.16)
Scored Live	0.23*	0.07	0.07	-0.13	-0.03	0.14	0.10	0.19*	0.33***	-0.12	0.33***
(β_{Live})	(0.09)	(0.05)	(0.10)	(0.09)	(0.07)	(0.08)	(0.10)	(0.10)	(0.09)	(0.10)	(0.08)
Double Scored	-0.13*	-0.01	-0.03	-0.08	0.02	0.01	0.00	-0.03	-0.05	-0.09	0.05
(β_{Dbl})	(0.06)	(0.03)	(0.07)	(0.06)	(0.04)	(0.05)	(0.06)	(0.06)	(0.06)	(0.07)	(0.05)
Date Scored (m)	-0.01	0.01**	-0.04***	-0.04***	0.01*	0.02***	-0.04***	-0.03***	-0.02**	-0.04***	0.00
(β_{DtSc})	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Day of the Week (β_{DayWk})											
Tuesday	0.01	0.03	0.00	0.00	0.10	0.02	0.00	-0.02	-0.02	-0.09	0.04
	(0.06)	(0.03)	(0.07)	(0.07)	(0.05)	(0.05)	(0.07)	(0.07)	(0.06)	(0.07)	(0.06)
Wednesday	0.11	0.05	0.17*	0.08	0.11*	0.05	0.12	0.08	0.07	0.14	0.17**
	(0.07)	(0.04)	(0.08)	(0.07)	(0.05)	(0.06)	(0.07)	(0.07)	(0.06)	(0.08)	(0.06)
Thursday	0.09	0.01	0.06	0.05	0.03	0.03	0.05	-0.11	-0.05	-0.02	0.04
	(0.07)	(0.03)	(0.08)	(0.07)	(0.05)	(0.06)	(0.07)	(0.07)	(0.06)	(0.07)	(0.06)
Friday	-0.08	-0.01	-0.04	-0.11	0.03	-0.00	-0.09	-0.17	-0.13	-0.17	0.04
	(0.08)	(0.04)	(0.09)	(0.08)	(0.06)	(0.07)	(0.08)	(0.09)	(0.07)	(0.09)	(0.07)
Observation	-0.03**	-0.02**	-0.03*	-0.02	-0.03***	-0.03**	-0.03**	-0.03*	-0.04***	-0.02	-0.03**
Month (β_{Month})	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Segment											
2	0.04	0.02	0.23***	0.10***	-0.04*	0.05*	0.06*	0.06*	0.17***	0.18***	0.03
	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)
3	0.02	0.01	0.27***	0.12***	-0.10***	0.03	-0.06	-0.08*	0.20***	0.16***	0.02
	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)
4+	0.03	0.01	0.34***	-0.02	-0.19***	-0.04	-0.23***	-0.28***	0.16***	0.11*	0.02
	(0.03)	(0.02)	(0.05)	(0.04)	(0.03)	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.03)
Content Domain											
Reading (β_{Read})	-0.00	0.01	-0.14	-0.10	0.08	0.18*	-0.02	0.03	0.04	-0.14	0.07
	(0.08)	(0.04)	(0.10)	(0.09)	(0.06)	(0.07)	(0.09)	(0.09)	(0.08)	(0.09)	(0.07)
Literature (β_{Lit})	0.09	0.05	0.20**	0.04	0.01	0.09	0.14*	0.23***	0.17**	0.27***	0.07
	(0.06)	(0.03)	(0.07)	(0.06)	(0.05)	(0.05)	(0.06)	(0.07)	(0.05)	(0.07)	(0.05)
Writing (β_{Write})	0.09	0.06*	0.04	0.16**	0.06	0.08	0.13*	0.13*	0.20***	0.24***	0.07
	(0.06)	(0.03)	(0.07)	(0.06)	(0.04)	(0.05)	(0.06)	(0.06)	(0.05)	(0.06)	(0.05)
Grammar	0.05	-0.00	-0.11	0.04	-0.02	-0.00	0.08	0.24***	-0.07	0.18**	-0.01
($\beta_{Grammar}$)	(0.06)	(0.03)	(0.07)	(0.06)	(0.04)	(0.05)	(0.06)	(0.06)	(0.05)	(0.06)	(0.05)
Interaction Structure											
Discussion	0.09*	0.05*	0.22***	0.08	0.00	0.02	0.14**	0.17***	0.06	0.13*	0.04
(β_{Disc})	(0.04)	(0.02)	(0.05)	(0.05)	(0.03)	(0.04)	(0.05)	(0.05)	(0.04)	(0.05)	(0.04)
Independent	0.03	-0.03	0.02	0.09	0.04	0.02	0.04	0.07	-0.04	-0.04	0.01
(β_{Ind})	(0.07)	(0.04)	(0.08)	(0.07)	(0.06)	(0.06)	(0.08)	(0.08)	(0.07)	(0.08)	(0.06)
Recitation	-0.02	-0.05	-0.07	-0.07	-0.02	-0.06	-0.05	-0.02	-0.05	-0.01	-0.06
(β_{Rec})	(0.05)	(0.03)	(0.06)	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)	(0.05)	(0.06)	(0.04)

Note. Each column shows the results of a separate model for the indicated item and instrument. Date Scored is scaled so a 1 point difference is one month; Monday is the reference group for the Days of the Week. PC=Positive Climate; NC=Negative Climate; RSP=Regard for Adolescent Behavior; TS=Teacher Sensitivity; BM=Behavior Management; PD=Productivity; ILF=Instructional Learning Formats; CU=Content Understanding; APS=Analysis and Problem Solving; QF=Quality of Feedback; ENG=Student Engagement.

Negative Climate has been reverse coded so higher scores capture higher quality. * p<0.05; ** p<0.01; *** p<0.001.

Table D.11: Item-Specific Fixed Effects from CI GTheory Model for Instrument FFT in Scale Score Metric

Facet	RR	CL	MCP	MSB	OPS	CS	KC	QDT	ESL	UAI	FR
Intercept	2.85***	2.39***	2.52***	2.88***	2.32***	2.63***	2.15***	1.97***	2.30***	1.92***	2.16***
	(0.05)	(0.07)	(0.08)	(0.06)	(0.06)	(0.07)	(0.08)	(0.06)	(0.09)	(0.08)	(0.07)
Scored Live	0.09	0.29***	-0.02	-0.06	-0.11	-0.04	0.21**	0.19**	0.13*	-0.03	0.23***
(β_{Live})	(0.05)	(0.07)	(0.07)	(0.05)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)
Double Scored	-0.01	-0.00	-0.06	0.00	-0.00	0.02	0.03	0.04	0.02	0.07	0.03
(β_{Dbl})	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Date Scored (m)	0.00	-0.00	-0.01	-0.00	-0.02***	-0.01***	-0.01**	0.00	-0.00	-0.01***	-0.01*
(β_{DtSc})	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Day of the Week (β_{DayWk})											
Tuesday	0.00	0.03	0.03	0.01	0.02	0.06	-0.01	-0.01	0.03	-0.00	-0.00
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Wednesday	0.08*	0.11*	0.07	0.05	0.05	0.07	0.06	0.04	0.09	0.07	0.05
	(0.04)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)
Thursday	0.05	0.11*	0.02	0.02	0.09*	0.01	0.03	0.00	0.06	0.04	0.02
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)
Friday	0.01	-0.03	-0.01	-0.01	-0.02	-0.03	-0.04	-0.03	-0.03	-0.06	0.01
	(0.04)	(0.05)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)
Observation	-0.02**	-0.03***	-0.01*	-0.02**	-0.01	-0.02*	-0.01	-0.02**	-0.02*	0.00	-0.02**
Month (β_{Month})	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Segment 2	-0.00	-0.03	-0.03	-0.07*	0.01	-0.08*	-0.10**	-0.03	-0.06	0.01	0.00
	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)
Content Domain											
Reading (β_{Read})	-0.06	-0.10	0.01	0.06	-0.04	-0.02	-0.05	-0.13*	-0.15**	-0.11*	-0.01
	(0.04)	(0.05)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)
Literature (β_{Lit})	0.05	0.09*	0.08	0.05	0.04	0.13***	0.12**	0.13***	0.17***	-0.01	0.03
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)
Writing (β_{Write})	0.06*	-0.03	0.04	-0.01	0.01	0.02	0.03	-0.04	0.05	0.13***	0.06
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)
Grammar	-0.04	-0.09*	-0.05	-0.02	-0.05	-0.07	-0.04	-0.09**	-0.07	-0.01	-0.05
($\beta_{Grammar}$)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)
Interaction Structure											
Discussion	0.02	-0.02	0.02	-0.03	0.07*	-0.00	0.01	0.10***	0.01	0.02	0.00
(β_{Disc})	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Independent	-0.00	0.07	0.07	-0.03	0.08	-0.02	0.11*	0.08	0.07	0.02	-0.01
(β_{Ind})	(0.04)	(0.05)	(0.05)	(0.04)	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)	(0.04)	(0.04)
Recitation	0.01	0.02	-0.01	-0.01	-0.04	0.04	0.05	0.03	-0.05	0.02	0.01
(β_{Rec})	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)

Note. Each column shows the results of a separate model for the indicated item and instrument. Date Scored is scaled so a 1 point difference is one month; Monday is the reference group for the Days of the Week. RR=Respect and Rapport; CL=Culture for Learning; MCP=Managing Classroom Procedures; MSB=Managing Student Behavior; OPS=Organizing Physical Space; CS=Communicating with Students; KC=Knowledge of Content and Pedagogy; QDT=Questioning Discussion Techniques; ESL=Engaging Students in Learning; UAI=Using Assessment in Instruction; FR=Flexibility and Responsiveness. * p<0.05; ** p<0.01; *** p<0.001.

Table D.12: Item-Specific Fixed Effects from CI GTheory Model for Instrument PLATO in Scale Score Metric

Facet	PURP	INTC	RC	CPK	CPE	ESI	MOD	GP	CD	TBI	ALL	BMN	TMN
Intercept	2.73***	1.96***	2.19***	1.58***	1.27***	1.11***	1.16***	2.33***	1.92***	1.42***	1.24***	3.92***	3.71***
	(0.07)	(0.10)	(0.07)	(0.08)	(0.07)	(0.05)	(0.06)	(0.15)	(0.09)	(0.11)	(0.07)	(0.03)	(0.06)
Scored Live	0.10*	-0.05	-0.07	-0.09	-0.16**	0.10*	0.13**	-0.23**	-0.24***	-0.13	0.07	0.02	-0.01
(β_{Live})	(0.05)	(0.05)	(0.06)	(0.06)	(0.05)	(0.04)	(0.05)	(0.08)	(0.06)	(0.10)	(0.05)	(0.03)	(0.05)
Double Scored	-0.07*	-0.07*	-0.11**	-0.02	0.00	-0.03	-0.07*	-0.07	-0.05	-0.09	-0.06	-0.00	0.01
(β_{Dbl})	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.02)	(0.03)
Date Scored	0.00	-0.01**	0.00	-0.01***	-0.02***	-0.01*	-0.01	-0.00	-0.03***	-0.01	-0.02***	0.00*	0.01**
(m) (β_{DtSc})	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Day of the Week (β_{DayWk})													
Tuesday	0.03	0.01	0.01	0.01	0.04	-0.01	-0.06	0.03	-0.03	-0.04	-0.05	0.04	0.02
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.05)	(0.04)	(0.06)	(0.03)	(0.02)	(0.03)
Wednesday	0.06	0.06	0.01	-0.00	0.05	-0.04	-0.08*	0.08	0.03	0.08	-0.06	0.02	0.02
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.06)	(0.04)	(0.07)	(0.03)	(0.02)	(0.03)
Thursday	-0.03	-0.00	-0.03	-0.02	0.03	-0.03	-0.06	-0.04	-0.02	0.05	-0.04	0.00	-0.01
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.05)	(0.04)	(0.06)	(0.03)	(0.02)	(0.03)
Friday	0.01	0.02	0.04	0.01	0.04	-0.06	-0.07	0.00	-0.08	-0.02	-0.02	-0.02	0.02
	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.03)	(0.04)	(0.07)	(0.05)	(0.08)	(0.04)	(0.03)	(0.04)
Observation Month	-0.00	-0.02**	-0.01	-0.01	-0.00	-0.00	-0.01	-0.01	-0.01*	0.00	-0.00	-0.01*	-0.02***
(β_{Month})	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Segment													
2	0.01	0.12***	0.08***	-0.16***	0.05*	0.02	0.10***	0.13***	0.14***	0.29***	0.02	-0.03**	0.10***
	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)
3	0.00	0.15***	0.01	-0.31***	-0.01	-0.02	0.08***	0.22***	0.12***	0.32***	-0.02	-0.04**	0.10***
	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)
4+	-0.03	0.12***	-0.12**	-0.45***	-0.04	-0.04	0.03	0.20***	0.10***	0.34***	-0.09***	-0.04*	0.10***
	(0.02)	(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.04)	(0.03)	(0.05)	(0.03)	(0.02)	(0.03)
Content Domain													
Reading	-0.01	0.04	-0.01	-0.00	0.04	0.08*	-0.03	-0.10	-0.05	0.52***	0.09*	0.03	0.03
(β_{Read})	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.03)	(0.04)	(0.07)	(0.05)	(0.08)	(0.04)	(0.03)	(0.04)
Literature	0.08**	0.14***	0.16***	0.13***	0.09*	-0.02	-0.02	0.03	0.18***	0.64***	0.07*	0.02	0.05
(β_{Lit})	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.02)	(0.03)	(0.05)	(0.04)	(0.06)	(0.03)	(0.02)	(0.03)
Writing	0.15***	0.20***	0.20***	0.02	0.01	0.06*	0.18***	0.23***	-0.01	0.18**	0.04	0.03	0.06*
(β_{Write})	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.02)	(0.03)
Grammar	0.04	0.03	0.17***	0.02	-0.13***	0.04	-0.01	0.08	-0.07*	-0.27***	0.10***	-0.00	-0.00
($\beta_{Grammar}$)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.02)	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.02)	(0.03)
Interaction Structure													
Discussion	0.10***	0.12***	0.11***	0.06*	0.06*	0.04*	0.03	0.12**	0.20***	0.11*	0.03	0.02	0.03
(β_{Disc})	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.04)	(0.03)	(0.05)	(0.02)	(0.02)	(0.02)
Independent	0.10**	0.02	0.12*	0.02	0.05	-0.00	-0.02	0.11	0.07	0.04	0.04	0.02	0.06
(β_{Ind})	(0.03)	(0.04)	(0.05)	(0.05)	(0.04)	(0.03)	(0.04)	(0.06)	(0.04)	(0.07)	(0.04)	(0.03)	(0.04)
Recitation	-0.01	-0.02	0.10**	0.05	-0.03	0.01	0.03	-0.09	0.01	0.10	0.06*	0.02	0.05
(β_{Rec})	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.02)	(0.03)

Note. Each column shows the results of a separate model for the indicated item and instrument. Date Scored is scaled so a 1 point difference is one month; Monday is the reference group for the Days of the Week. PURP=Purpose; INTC=Intellectual Climate; RC=Representation of Content; CPK=Connections to Prior Knowledge; CPE=Connections to Personal and /or Cultural Experience; ESI=Explicit Strategy Instruction; MOD=Modeling; GP=Guided Practice; CD=Classroom Discussion; TBI=Text-Based Instruction; ALL=Accommodations for Language Learners; BMN=Behavior Management; TMN=Time Management. * p<0.05; ** p<0.01; *** p<0.001.

Table D.13: Item-Specific Fixed Effects from SO GTheory Model for Instrument CLASS in Scale Score Metric

Facet	PC	NC	RSP	TS	BM	PD	ILF	CU	APS	QF	ENG
Intercept	4.68***	6.83***	2.94***	4.02***	6.17***	5.86***	3.74***	3.20***	2.37***	3.19***	5.13***
	(0.21)	(0.09)	(0.19)	(0.21)	(0.11)	(0.14)	(0.16)	(0.16)	(0.17)	(0.20)	(0.16)
Scored Live (β_{Live})	0.28**	0.08	0.12	-0.09	-0.01	0.17*	0.14	0.23*	0.37***	-0.07	0.38***
	(0.09)	(0.05)	(0.10)	(0.09)	(0.07)	(0.08)	(0.10)	(0.10)	(0.08)	(0.10)	(0.08)
Double Scored	-0.15*	-0.02	-0.05	-0.09	0.01	-0.00	-0.02	-0.05	-0.06	-0.11	0.03
(β_{Dbl})	(0.06)	(0.03)	(0.07)	(0.06)	(0.04)	(0.05)	(0.06)	(0.06)	(0.06)	(0.07)	(0.05)
Date Scored (m)	-0.01	0.01***	-0.04***	-0.04***	0.01**	0.02***	-0.03***	-0.02**	-0.01*	-0.03***	0.01
(β_{DtSc})	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Day of the Week (β_{DayWk})											
Tuesday	-0.00	0.02	-0.01	-0.01	0.08	0.01	-0.01	-0.04	-0.03	-0.10	0.03
	(0.06)	(0.03)	(0.07)	(0.06)	(0.05)	(0.05)	(0.06)	(0.07)	(0.06)	(0.07)	(0.06)
Wednesday	0.07	0.04	0.13	0.04	0.08	0.03	0.09	0.04	0.04	0.11	0.14*
	(0.07)	(0.04)	(0.08)	(0.07)	(0.05)	(0.06)	(0.07)	(0.07)	(0.06)	(0.07)	(0.06)
Thursday	0.07	0.00	0.04	0.03	0.01	0.02	0.04	-0.13	-0.07	-0.03	0.03
	(0.06)	(0.03)	(0.07)	(0.07)	(0.05)	(0.05)	(0.07)	(0.07)	(0.06)	(0.07)	(0.06)
Friday	-0.07	-0.01	-0.05	-0.10	0.03	-0.00	-0.09	-0.17	-0.13	-0.17	0.04
	(0.08)	(0.04)	(0.09)	(0.08)	(0.06)	(0.07)	(0.08)	(0.09)	(0.07)	(0.09)	(0.07)
Observation Month	-0.03**	-0.02**	-0.03**	-0.02	-0.03***	-0.03**	-0.03**	-0.03*	-0.04***	-0.02*	-0.03***
(β_{Month})	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Segment											
2	0.04	0.02	0.24***	0.10***	-0.04*	0.05*	0.06*	0.06*	0.17***	0.18***	0.04
	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)
3	0.02	0.01	0.27***	0.12***	-0.10***	0.03	-0.06	-0.08*	0.20***	0.16***	0.02
	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)
4+	0.03	0.02	0.35***	-0.01	-0.19***	-0.03	-0.22***	-0.27***	0.16***	0.12*	0.03
	(0.03)	(0.02)	(0.05)	(0.04)	(0.03)	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.03)
Content Domain											
Reading (β_{Read})	0.04	0.02	-0.10	-0.06	0.10	0.20**	0.02	0.06	0.06	-0.10	0.10
	(0.08)	(0.04)	(0.09)	(0.08)	(0.06)	(0.07)	(0.08)	(0.09)	(0.07)	(0.09)	(0.07)
Literature (β_{Lit})	0.04	0.03	0.13	-0.02	-0.02	0.06	0.10	0.18**	0.12*	0.21**	0.03
	(0.06)	(0.03)	(0.07)	(0.06)	(0.05)	(0.05)	(0.06)	(0.06)	(0.05)	(0.06)	(0.05)
Writing (β_{Write})	0.09	0.06*	0.03	0.16**	0.05	0.08	0.14*	0.13*	0.20***	0.25***	0.07
	(0.05)	(0.03)	(0.06)	(0.06)	(0.04)	(0.05)	(0.06)	(0.06)	(0.05)	(0.06)	(0.05)
Grammar	0.05	0.00	-0.11	0.05	-0.00	0.01	0.08	0.24***	-0.06	0.18**	-0.00
($\beta_{Grammar}$)	(0.06)	(0.03)	(0.07)	(0.06)	(0.04)	(0.05)	(0.06)	(0.06)	(0.05)	(0.06)	(0.05)
Interaction Structure											
Discussion (β_{Disc})	0.05	0.03	0.19***	0.05	-0.01	0.00	0.11*	0.15**	0.04	0.10*	0.01
	(0.04)	(0.02)	(0.05)	(0.04)	(0.03)	(0.04)	(0.05)	(0.05)	(0.04)	(0.05)	(0.04)
Independent	0.03	-0.03	0.02	0.08	0.03	0.01	0.04	0.06	-0.05	-0.04	0.01
(β_{Ind})	(0.07)	(0.04)	(0.08)	(0.07)	(0.06)	(0.06)	(0.07)	(0.08)	(0.07)	(0.08)	(0.06)
Recitation (β_{Rec})	-0.02	-0.05*	-0.07	-0.06	-0.02	-0.07	-0.05	-0.02	-0.05	-0.00	-0.06
	(0.05)	(0.03)	(0.06)	(0.05)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)
Grade											
7th Grade (β_{7th})	-0.21**	-0.10**	-0.13	-0.19**	-0.09	-0.11*	-0.25***	-0.12	-0.05	-0.16*	-0.16**
	(0.07)	(0.04)	(0.07)	(0.07)	(0.05)	(0.05)	(0.07)	(0.07)	(0.05)	(0.07)	(0.06)
8th Grade (β_{8th})	-0.06	-0.03	0.05	0.02	0.08	0.08	-0.05	0.02	0.08	0.00	0.01
	(0.06)	(0.03)	(0.07)	(0.07)	(0.05)	(0.05)	(0.06)	(0.07)	(0.05)	(0.06)	(0.05)
Prior Achievement	0.09*	0.03	0.14**	0.08	0.09**	0.07*	0.08*	0.06	0.06*	0.11**	0.12***
(β_{PrAch})	(0.04)	(0.02)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)
St. Info Missing	-0.18	-0.01	-0.18	-0.12	-0.09	-0.17*	-0.14	-0.15	-0.06	-0.13	-0.18
(β_{Imp})	(0.11)	(0.06)	(0.12)	(0.11)	(0.08)	(0.08)	(0.11)	(0.11)	(0.09)	(0.11)	(0.09)
Demographic	-0.18***	-0.04	-0.10*	-0.15***	-0.03	-0.03	-0.11**	-0.10*	-0.08*	-0.08*	-0.07*
Composite (β_{Demo})	(0.04)	(0.02)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)

Note. Each column shows the results of a separate model for the indicated item and instrument. Date Scored is scaled so a 1 point difference is one month. Monday is the reference group for the Days of the Week. The Demographic Composite represents classrooms that have higher percentages of students who are black, Hispanic, ELL, and FRL. St. Info Missing is a dummy variable indicating if Prior Achievement and Demographic Composite are missing. PC=Positive Climate; NC=Negative Climate; RSP=Regard for Adolescent Behavior; TS=Teacher Sensitivity; BM=Behavior Management; PD=Productivity;

ILF=Instructional Learning Formats; CU=Content Understanding; APS=Analysis and Problem Solving; QF=Quality of Feedback; ENG=Student Engagement. Negative Climate has been reverse coded so higher scores capture higher quality. * p<0.05; ** p<0.01; *** p<0.001.

Table D.14: Item-Specific Fixed Effects from SO GTheory Model for Instrument FFT in Scale Score Metric

Facet	RR	CL	MCP	MSB	OPS	CS	KC	QDT	ESL	UAI	FR
Intercept	2.90***	2.48***	2.53***	2.89***	2.38***	2.70***	2.21***	2.02***	2.37***	1.95***	2.22***
	(0.06)	(0.08)	(0.08)	(0.06)	(0.06)	(0.08)	(0.08)	(0.07)	(0.09)	(0.08)	(0.07)
Scored Live (β_{Live})	0.11*	0.33***	0.00	-0.05	-0.08	-0.01	0.25***	0.22***	0.17*	-0.01	0.27***
	(0.05)	(0.06)	(0.07)	(0.05)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Double Scored	-0.01	-0.01	-0.07	-0.00	-0.02	0.02	0.02	0.03	0.00	0.06	0.02
(β_{Dbl})	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Date Scored (m)	0.01*	0.00	-0.00	-0.00	-0.02***	-0.01**	-0.01*	0.00	-0.00	-0.01**	-0.01
(β_{DtSc})	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Day of the Week (β_{DayWk})											
Tuesday	-0.01	0.01	0.02	0.00	-0.00	0.04	-0.03	-0.03	0.02	-0.02	-0.02
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Wednesday	0.06	0.10*	0.05	0.04	0.02	0.05	0.03	0.02	0.07	0.05	0.03
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)
Thursday	0.03	0.09*	0.01	0.01	0.07	-0.02	-0.00	-0.02	0.04	0.01	-0.01
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Friday	-0.00	-0.04	-0.01	-0.01	-0.05	-0.05	-0.06	-0.04	-0.04	-0.07	-0.01
	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Observation Month	-0.02**	-0.03***	-0.02*	-0.02***	-0.01	-0.02*	-0.01	-0.02**	-0.02*	0.00	-0.02**
(β_{Month})	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Segment 2	0.00	-0.01	-0.01	-0.06*	0.02	-0.07*	-0.09**	-0.02	-0.05	0.02	0.02
	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)
Content Domain											
Reading (β_{Read})	-0.04	-0.08	0.03	0.08	-0.01	-0.01	-0.02	-0.10*	-0.12*	-0.08	0.02
	(0.04)	(0.05)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Literature (β_{Lit})	0.03	0.05	0.06	0.03	0.01	0.10**	0.07	0.10**	0.13**	-0.04	-0.01
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)
Writing (β_{Write})	0.06	-0.03	0.03	-0.01	0.00	0.02	0.03	-0.05	0.05	0.13***	0.05
	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
Grammar	-0.04	-0.09*	-0.03	-0.01	-0.04	-0.07	-0.04	-0.09**	-0.07	0.00	-0.05
($\beta_{Grammar}$)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
Interaction Structure											
Discussion (β_{Disc})	0.01	-0.04	0.00	-0.04	0.05	-0.02	-0.01	0.08**	-0.01	-0.00	-0.02
	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Independent	-0.00	0.06	0.06	-0.03	0.08	-0.02	0.10*	0.07	0.05	0.02	-0.02
(β_{Ind})	(0.04)	(0.05)	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)
Recitation (β_{Rec})	0.01	0.03	-0.01	-0.01	-0.04	0.03	0.05	0.03	-0.05	0.02	0.01
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Grade											
7th Grade (β_{7th})	-0.05	-0.12**	-0.07	-0.04	-0.07	-0.07	-0.06	-0.04	-0.09*	-0.00	-0.06
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)
8th Grade (β_{8th})	0.01	-0.05	0.08	0.07	-0.00	-0.01	0.00	0.01	-0.03	0.02	0.02
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)
Prior Achievement	0.03	0.13***	0.10***	0.09***	0.06*	0.03	0.12***	0.10***	0.12***	0.06**	0.07***
(β_{PrAch})	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
St. Info Missing	-0.11	-0.03	-0.05	-0.06	-0.08	-0.10	-0.03	-0.04	-0.03	-0.06	-0.06
(β_{Imp})	(0.06)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.05)	(0.05)
Demographic	-0.05*	-0.03	0.01	0.00	-0.04	-0.07**	-0.03	-0.02	-0.02	-0.02	-0.04*
Composite (β_{Demo})	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)

Note. Each column shows the results of a separate model for the indicated item and instrument. Date Scored is scaled so a 1 point difference is one month. Monday is the reference group for the Days of the Week. The Demographic Composite represents classrooms that have higher percentages of students who are black, Hispanic, ELL, and FRL. St. Info Missing is a dummy variable indicating if Prior Achievement and Demographic Composite are missing. RR=Respect and Rapport; CL=Culture for Learning; MCP=Managing Classroom Procedures; MSB=Managing Student Behavior; OPS=Organizing Physical Space; CS=Communicating with Students; KC=Knowledge of Content and Pedagogy; QDT=Questioning Discussion Techniques; ESL=Engaging Students in Learning; UAI=Using Assessment in Instruction; FR=Flexibility and Responsiveness. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

Table D.15: Item-Specific Fixed Effects from SO GTheory Model for Instrument PLATO in Scale Score Metric

Facet	PURP	INTC	RC	CPK	CPE	ESI	MOD	GP	CD	TBI	ALL	BMN	TMN
Intercept	2.76***	2.00***	2.23***	1.59***	1.29***	1.11***	1.15***	2.40***	1.99***	1.43***	1.26***	3.92***	3.74***
	(0.07)	(0.10)	(0.08)	(0.08)	(0.08)	(0.06)	(0.06)	(0.15)	(0.09)	(0.12)	(0.07)	(0.04)	(0.06)
Scored Live	0.11*	-0.02	-0.04	-0.08	-0.15**	0.09*	0.12*	-0.21*	-0.20***	-0.10	0.07	0.03	0.00
(β_{Live})	(0.05)	(0.05)	(0.06)	(0.06)	(0.05)	(0.04)	(0.05)	(0.08)	(0.06)	(0.10)	(0.05)	(0.03)	(0.05)
Double Scored	-0.07**	-0.08*	-0.12***	-0.02	0.00	-0.03	-0.07*	-0.07	-0.06	-0.10	-0.06	-0.00	0.01
(β_{Dbl})	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.02)	(0.03)
Date Scored (m)	0.00	-0.01*	0.00	-0.01***	-0.01***	-0.01*	-0.01	-0.00	-0.02***	-0.00	-0.02***	0.01*	0.01***
(β_{DtSc})	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Day of the Week (β_{DayWk})													
Tuesday	0.03	-0.00	-0.00	0.00	0.03	-0.01	-0.06	0.02	-0.05	-0.05	-0.05	0.03	0.01
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.05)	(0.04)	(0.06)	(0.03)	(0.02)	(0.03)
Wednesday	0.05	0.04	-0.01	-0.01	0.04	-0.03	-0.08*	0.06	0.01	0.06	-0.05	0.01	0.01
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.06)	(0.04)	(0.07)	(0.03)	(0.02)	(0.03)
Thursday	-0.04	-0.02	-0.05	-0.02	0.02	-0.03	-0.06	-0.06	-0.03	0.03	-0.04	-0.01	-0.03
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.05)	(0.04)	(0.06)	(0.03)	(0.02)	(0.03)
Friday	0.01	0.01	0.02	0.01	0.03	-0.06	-0.07	-0.01	-0.08	-0.02	-0.03	-0.03	0.01
	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.03)	(0.04)	(0.06)	(0.05)	(0.08)	(0.04)	(0.03)	(0.04)
Observation	-0.00	-0.02**	-0.01	-0.01	-0.00	-0.00	-0.01	-0.01	-0.01*	-0.00	-0.00	-0.01*	-0.02***
Month (β_{Month})	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Segment													
2	0.01	0.12***	0.08***	-0.16***	0.05*	0.02	0.10***	0.13***	0.14***	0.29***	0.02	-0.03**	0.10***
	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)
3	0.00	0.15***	0.01	-0.31***	-0.01	-0.02	0.08***	0.22***	0.12***	0.32***	-0.02	-0.04**	0.10***
	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)
4+	-0.02	0.13***	-0.10**	-0.44***	-0.04	-0.04*	0.03	0.21***	0.11***	0.34***	-0.09***	-0.04*	0.11***
	(0.02)	(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.04)	(0.03)	(0.05)	(0.03)	(0.02)	(0.03)
Content Domain													
Reading	-0.00	0.05	0.01	0.00	0.05	0.08*	-0.03	-0.09	-0.03	0.54***	0.09*	0.04	0.04
(β_{Read})	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.03)	(0.04)	(0.07)	(0.05)	(0.08)	(0.04)	(0.03)	(0.04)
Literature	0.07*	0.11***	0.13***	0.12**	0.07	-0.02	-0.01	0.01	0.14***	0.61***	0.07*	0.00	0.03
(β_{Lit})	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.02)	(0.03)
Writing	0.15***	0.20***	0.19***	0.02	0.01	0.06*	0.18***	0.23***	-0.01	0.17**	0.05	0.02	0.05*
(β_{Write})	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.02)	(0.03)
Grammar	0.04	0.03	0.17***	0.02	-0.13***	0.05*	-0.01	0.08	-0.07*	-0.25***	0.09***	0.00	0.00
($\beta_{Grammar}$)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.02)	(0.03)
Interaction Structure													
Discussion	0.09***	0.11***	0.10**	0.05	0.05	0.05*	0.04	0.11**	0.19***	0.10*	0.03	0.02	0.02
(β_{Disc})	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.04)	(0.03)	(0.05)	(0.02)	(0.02)	(0.02)
Independent	0.10**	0.02	0.12*	0.03	0.04	-0.01	-0.02	0.11	0.07	0.04	0.04	0.02	0.06
(β_{Ind})	(0.03)	(0.04)	(0.05)	(0.05)	(0.04)	(0.03)	(0.04)	(0.06)	(0.04)	(0.07)	(0.04)	(0.03)	(0.04)
Recitation	-0.02	-0.02	0.10**	0.05	-0.03	0.01	0.03	-0.10*	0.01	0.10	0.06*	0.02	0.05
(β_{Rec})	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.02)	(0.03)
Grade													
7th Grade (β_{7th})	-0.03	-0.06	-0.05	0.00	-0.04	-0.05	-0.01	-0.12*	-0.09*	-0.03	-0.06*	0.01	-0.04
	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.05)	(0.04)	(0.06)	(0.03)	(0.02)	(0.03)
8th Grade (β_{8th})	-0.01	0.01	0.04	-0.01	0.01	0.01	0.03	0.01	-0.02	0.08	-0.03	0.04	0.03
	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.02)	(0.03)
Prior Achieve	-0.00	0.04*	0.01	-0.00	-0.00	0.01	-0.03	-0.01	0.08***	0.11**	-0.02	0.03*	0.03
(β_{PrAch})	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.04)	(0.02)	(0.01)	(0.02)
St. Info Missing	-0.10*	-0.07	-0.11	-0.00	-0.00	0.04	-0.02	-0.12	-0.03	-0.07	0.11*	-0.03	-0.09*
(β_{Imp})	(0.04)	(0.05)	(0.06)	(0.06)	(0.05)	(0.04)	(0.05)	(0.07)	(0.06)	(0.10)	(0.05)	(0.04)	(0.05)
Demographic Composite	-0.02	-0.04*	-0.09***	-0.03	-0.04*	0.01	-0.02	-0.06*	-0.04	0.01	-0.03	-0.01	-0.02
(β_{Demo})	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.04)	(0.02)	(0.01)	(0.02)

Note. Each column shows the results of a separate model for the indicated item and instrument. Date Scored is scaled so a 1 point difference is one month. Monday is the reference group for the Days of the Week. The Demographic Composite represents classrooms that have higher percentages of students who are black, Hispanic, ELL, and FRL. St. Info Missing is a dummy variable indicating if Prior Achievement and Demographic Composite are missing. PURP=Purpose; INTC=Intellectual Climate; RC=Representation of Content; CPK=Connections to Prior Knowledge; CPE=Connections to Personal and /or Cultural Experience; ESI=Explicit Strategy Instruction; MOD=Modeling; GP=Guided Practice; CD=Classroom Discussion; TBI=Text-Based Instruction; ALL=Accommodations for Language Learners; BMN=Behavior Management; TMN=Time Management. * p<0.05; ** p<0.01; *** p<0.001.

Appendix E – Sensitivity Analyses of Creation of Hidden Facets

This section provides a comparison of an alternative method of creating the Content Domain and Interaction Structure variables. I initially created these variables to highlight whether a day of instruction had a sustained focus on the relevant content domain or interaction structure. This focused the analysis on days where the content domain or interaction structure was a prominent part of the lesson. An alternative way of creating these variables is to simply average PLATO log scores up to the lesson level. Note that the use of 30 minute occasions for FFT makes it impossible to use segment level variables in a consistent way across all instruments. Tables E.1 and E.2 show a comparison of the fixed effects of the CI and SO models across the two models. The Sustained Focus columns are the original models while the Average Value columns are the new models. There are some meaningful differences across the two models. CLASS and FFT both score Reading lessons lower than non-reading lessons only in the average value model, which creates a larger contrast to the positive effect on PLATO than in the sustained focus. Similarly, CLASS and PLATO score grammar lessons more positively than non-grammar lessons when using the average PLATO log values, which again highlights the contrast of two instruments with the third. The impact of adding SO facets into the equation also differs slightly across the two methods of capturing the CI facets as the literature effect on CLASS decreased more when using the sustained focus approach while the literature effect on FFT decreased more when using the average value approach. Using average PLATO log values creates a greater sense of instrument bias due to discrepant content domain effects across instruments than in the original models. On the other hand, the average value model leads FFT to have the same positive effect, though admittedly much smaller, for discussion lessons as the other two instruments. This effect is only in the CI model and not the SO model. Last, recitation lessons receive higher scores on PLATO than non-recitation lessons only for the model using average scores on the PLATO log, leaving PLATO as the only instrument with a positive view of recitation lessons. The broad patterns here are consistent, CI facets have large effects and those effects are not always consistent across the instruments. The differences, though, do suggest a need to better capture these variables. Using occasion level variables would help accomplish this, but only if the occasions capture the content domain and interaction structures under study. This requires a more careful separation of days of instruction into meaningful occasions.

Table E.1: Comparison of Sustained Focus and Average Value approaches to constructing CI facets for CI Model in Teacher SD Metric

	Sustained Focus CI Facets			Average Value CI Facets		
Names	CLASS	FFT	PLATO	CLASS	FFT	PLATO
Scored Live(β_{Live})	0.17 (0.09)	0.25 (0.09)**	-0.09 (0.05)	0.14 (0.09)	0.23 (0.09)*	-0.09 (0.05)
Double Scored(β_{Dbl})	-0.05 (0.06)	0.03 (0.06)	-0.09 (0.03)**	-0.04 (0.06)	0.03 (0.06)	-0.05 (0.03)
Date Scored (m)(β_{DtSc})	-0.03 (0.01)***	-0.02 (0.01)**	-0.01 (0.00)***	-0.03 (0.01)***	-0.02 (0.01)**	-0.01 (0.00)***
Day of the Week (β_{DayWk})						
Tuesday	-0.01 (0.07)	0.03 (0.06)	-0.01 (0.04)	0.02 (0.07)	0.03 (0.06)	0.02 (0.04)
Wednesday	0.14 (0.07)*	0.14 (0.07)*	0.02 (0.04)	0.16 (0.07)*	0.14 (0.07)*	0.03 (0.04)
Thursday	-0.00 (0.07)	0.09 (0.07)	-0.04 (0.04)	0.03 (0.07)	0.09 (0.07)	-0.01 (0.04)
Friday	-0.12 (0.08)	-0.07 (0.08)	-0.01 (0.05)	-0.07 (0.08)	-0.07 (0.08)	0.02 (0.04)
Observation Month (β_{Month})	-0.05 (0.01)***	-0.04 (0.01)***	-0.02 (0.01)*	-0.05 (0.01)***	-0.04 (0.01)***	-0.01 (0.01)*
Content Domain						
Reading (β_{Read})	0.04 (0.09)	-0.12 (0.08)	0.10 (0.05)*	-0.08 (0.04)*	-0.09 (0.04)*	0.04 (0.02)*
Literature (β_{Lit})	0.16 (0.06)**	0.20 (0.06)***	0.24 (0.03)***	0.48 (0.13)***	0.36 (0.12)**	0.66 (0.07)***
Writing (β_{Write})	0.19 (0.06)***	0.10 (0.06)	0.22 (0.03)***	0.10 (0.03)***	0.03 (0.03)	0.15 (0.01)***
Grammar ($\beta_{Grammar}$)	0.08 (0.06)	-0.11 (0.06)*	0.00 (0.03)	0.06 (0.03)*	-0.08 (0.03)**	0.05 (0.01)**
Interaction Structure						
Discussion (β_{Disc})	0.13 (0.05)**	0.01 (0.04)	0.15 (0.03)***	0.92 (0.18)***	0.36 (0.17)*	0.93 (0.10)***
Independent (β_{Ind})	0.02 (0.07)	0.10 (0.07)	0.10 (0.04)*	0.24 (0.20)	0.25 (0.20)	0.46 (0.12)***
Recitation (β_{Rec})	-0.08 (0.05)	-0.01 (0.05)	0.05 (0.03)	-0.03 (0.11)	0.01 (0.10)	0.20 (0.07)**

Note. Each column shows the results of a separate model for the indicated instrument. The left three columns show the hidden facets when estimated using the sustained focus approach that was used throughout this thesis. These columns match the results of Table 5.6. The right three columns average scores across the logs to form the same variables as a sensitivity analysis. Date Scored is scaled so a 1 point difference is one month. Monday is the reference group for the Days of the Week. Sixth grade is the references group for grade. The Demographic Composite represents classrooms that have higher percentages of students who are black, Hispanic, ELL, and FRL. St. Info Missing is a dummy variable indicating if Prior Achievement and Demographic Composite are missing. * p<0.05; ** p<0.01; *** p<0.001.

Table E.2: Comparison of Sustained Focus and Average Value approaches to constructing SO facets for SO Model in Teacher SD Metric

	Sustained Focus CI Facets			Average Value CI Facets		
Names	CLASS	FFT	PLATO	CLASS	FFT	PLATO
Scored Live (β_{Live})	0.22 (0.09)*	0.30 (0.09)**	-0.07 (0.05)	0.20 (0.09)*	0.28 (0.09)**	-0.07 (0.05)
Double Scored (β_{Dbl})	-0.07 (0.06)	0.01 (0.06)	-0.10 (0.03)**	-0.07 (0.06)	0.01 (0.06)	-0.06 (0.03)*
Date Scored (m) (β_{DtSc})	-0.02 (0.01)***	-0.01 (0.01)*	-0.01 (0.00)***	-0.02 (0.01)***	-0.01 (0.01)*	-0.01 (0.00)***
Day of the Week (β_{DayWk})						
Tuesday	-0.03 (0.06)	-0.01 (0.06)	-0.02 (0.04)	0.00 (0.06)	-0.01 (0.06)	0.01 (0.04)
Wednesday	0.11 (0.07)	0.11 (0.07)	0.01 (0.04)	0.12 (0.07)	0.11 (0.07)	0.01 (0.04)
Thursday	-0.02 (0.07)	0.05 (0.06)	-0.05 (0.04)	0.01 (0.06)	0.05 (0.06)	-0.03 (0.04)
Friday	-0.12 (0.08)	-0.09 (0.08)	-0.02 (0.05)	-0.08 (0.08)	-0.10 (0.08)	0.01 (0.04)
Observation Month (β_{Month})	-0.05 (0.01)***	-0.04 (0.01)***	-0.02 (0.01)**	-0.05 (0.01)***	-0.04 (0.01)***	-0.01 (0.01)*
Content Domain						
Reading (β_{Read})	0.07 (0.08)	-0.07 (0.08)	0.12 (0.05)*	-0.04 (0.04)	-0.04 (0.04)	0.07 (0.02)**
Literature (β_{Lit})	0.11 (0.06)	0.14 (0.06)*	0.21 (0.03)***	0.36 (0.13)**	0.21 (0.12)	0.60 (0.07)***
Writing (β_{Write})	0.19 (0.06)***	0.08 (0.05)	0.21 (0.03)***	0.10 (0.03)***	0.03 (0.03)	0.15 (0.01)***
Grammar ($\beta_{Grammar}$)	0.08 (0.06)	-0.11 (0.06)*	0.01 (0.03)	0.06 (0.03)*	-0.08 (0.03)**	0.05 (0.01)***
Interaction Structure						
Discussion (β_{Disc})	0.10 (0.04)*	-0.02 (0.04)	0.14 (0.03)***	0.81 (0.17)***	0.22 (0.17)	0.89 (0.10)***
Independent (β_{Ind})	0.02 (0.07)	0.09 (0.07)	0.10 (0.04)*	0.26 (0.19)	0.27 (0.19)	0.47 (0.12)***
Recitation (β_{Rec})	-0.07 (0.05)	-0.00 (0.05)	0.05 (0.03)	-0.03 (0.11)	0.02 (0.10)	0.19 (0.07)**
Grade						
7th Grade (β_{7th})	-0.23 (0.07)**	-0.17 (0.07)*	-0.09 (0.04)*	-0.23 (0.07)**	-0.17 (0.07)*	-0.10 (0.04)**
8th Grade (β_{8th})	0.03 (0.07)	0.03 (0.07)	0.03 (0.04)	0.05 (0.07)	0.03 (0.07)	0.03 (0.04)
Prior Achievement (β_{PrAch})	0.13 (0.04)**	0.20 (0.04)***	0.04 (0.02)	0.12 (0.04)**	0.20 (0.04)***	0.03 (0.02)
St. Info Missing (β_{Imp})	-0.21 (0.11)	-0.18 (0.11)	-0.07 (0.06)	-0.20 (0.11)	-0.16 (0.11)	-0.06 (0.06)
Demographic Composite (β_{Demo})	-0.15 (0.04)***	-0.09 (0.04)*	-0.06 (0.02)**	-0.14 (0.04)***	-0.09 (0.04)*	-0.06 (0.02)**

Note. Each column shows the results of a separate model for the indicated instrument. The left three columns show the hidden facets when estimated using the sustained focus approach that was used throughout this thesis. These columns match the results of Table 5.7. The right three columns average scores across the logs to form the same variables as a sensitivity analysis. Date Scored is scaled so a 1 point difference is one month. Monday is the reference group for the Days of the Week. Sixth grade is the references group for grade. The Demographic Composite represents classrooms that have higher percentages of students who are black, Hispanic, ELL, and FRL. St. Info Missing is a dummy variable indicating if Prior Achievement and Demographic Composite are missing. * p<0.05; ** p<0.01; *** p<0.001.

Appendix F – Bootstrap Instrument Bias Analysis

One of the questions raised in this thesis is the problem of instrument bias. An observation instrument may score specific types of lessons (e.g. lectures or discussions) systematically lower or higher than their true instructional quality. I explore this through looking at the hidden facet effect estimates, standardized to the teacher quality standard deviation metric, across the three instruments. Arguably, the teacher standard deviation metric provides a common metric across instruments because it represents the extent to which a teacher will move across the distribution of teacher quality as a result of being observed on a given facet. However, there are challenges to directly comparing the estimated regression parameters across models. The regressions were run on the same population so errors in the hidden facet regression parameters are likely correlated across the instruments. In order to get around this challenge, I make use of the bootstrap replicates to test for significant differences in the regression parameters. Each bootstrap replicate is generated from an independent, simulated sampled. This should reduce any relationship between the errors of the hidden facet parameter estimates. Note, though, that each of the simulated samples had the same distribution of hidden facets across teachers, days, and raters (e.g. 20% of teachers were observed on two days of literature in each simulated sample), which could, in principal, still cause some bias in the results of this analysis.

Under the assumption of no instrument bias, the estimated effects of a hidden facet are equivalent (except for sampling variation) across the instruments. This implies that the relationship between observed teaching quality and the hidden facet in each of the bootstrapped simulation samples is equivalent. Thus, looking at the estimates of the hidden facet effect across the bootstrapped samples, under the assumption of no instrument bias, we should see that the sampling variation of the effect within an instrument (across replications) is much larger than the differences of this effect across instruments. We can quantify this difference by estimating a bootstrapped p-value. I tested this by randomly selecting a hidden facet estimate from the bootstrapped replications for each of two instruments, for example CLASS and FFT. I then tested if the hidden facet effect estimate is larger on CLASS as compared to the estimate for FFT. Repeating this procedure 1,000 times, I then calculated the percentage of times that the CLASS estimate was larger than the FFT estimate, which gives the p-value for whether there is evidence of instrument bias for the given hidden facet (note that this is equivalent to the Mann–Whitney *U* test; Mann & Whitney, 1947). Because I conducted three comparisons for each hidden facet, I use a Bonferroni correction for the p-values, interpreting p-values below 0.0167 and above 0.983 as significant (i.e. 0.05/3 and 1-0.05/3).

Figures F.1-F.3 show the estimates of the hidden facet effects across the bootstrapped replications and across instruments. Figure F.1 shows the SD facets; Figure F.2 shows the CI facets (equivalent to Figure 5.1); and Figure F.3 shows the SO facets. In each graph, every small dot represents an estimate of the effect of that hidden facet on one of the bootstrapped replications. The boxes show the 95th percentile of the effect estimates and the line in the middle of the box shows the mean effect. Looking at Figure F.1, the top set of three boxes shows the estimates of the effect of being scored live for the PLATO, FFT, and CLASS instruments. Notice the large overlap between the distribution of the estimates for the FFT and CLASS scores, which is indicative of cases where there is no evidence of instrument bias ($p=0.211$). While the average estimated effect of live scoring on FFT scores was larger than the average effect on CLASS scores, the sampling variation of these effects was much larger than the difference between the two instruments. On the other hand, almost every bootstrapped estimate of the effect of live scoring on PLATO scores is lower than the estimated effects on FFT or CLASS scores. This is indicative of instrument bias, with the effect of live scoring on PLATO scores significantly lower than the effect on FFT ($p<0.001$)

or CLASS scores (p=0.008). I discuss the meaning and implications of these effects in the main body of the thesis. Here, I simply state that none of the other SD facet effects (aside from live scoring) in Figure F.1 show were statistically significant. Figure F.2 shows the results for the CI facets. The effects on PLATO scores are significantly larger than the effects on FFT scores on reading, literature, writing, and discussion lessons. The effects on PLATO scores were significantly larger than the effects on CLASS scores for literature, writing, and discussion lessons. The effect on CLASS scores was significantly larger than the effect on FFT scores only for grammar lessons. Figure F.3 shows the results for the SO facets, where there are no significant differences across instruments.

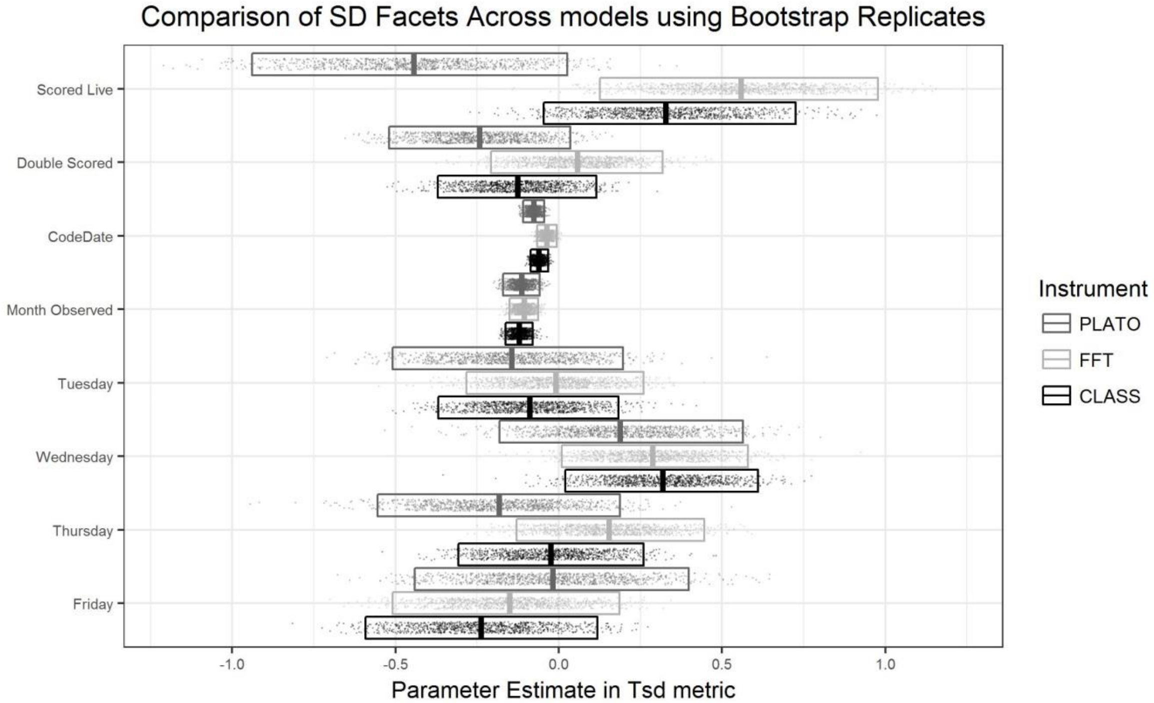

Figure F.1: Comparison of SD Facet Effects across Bootstrap Replicates

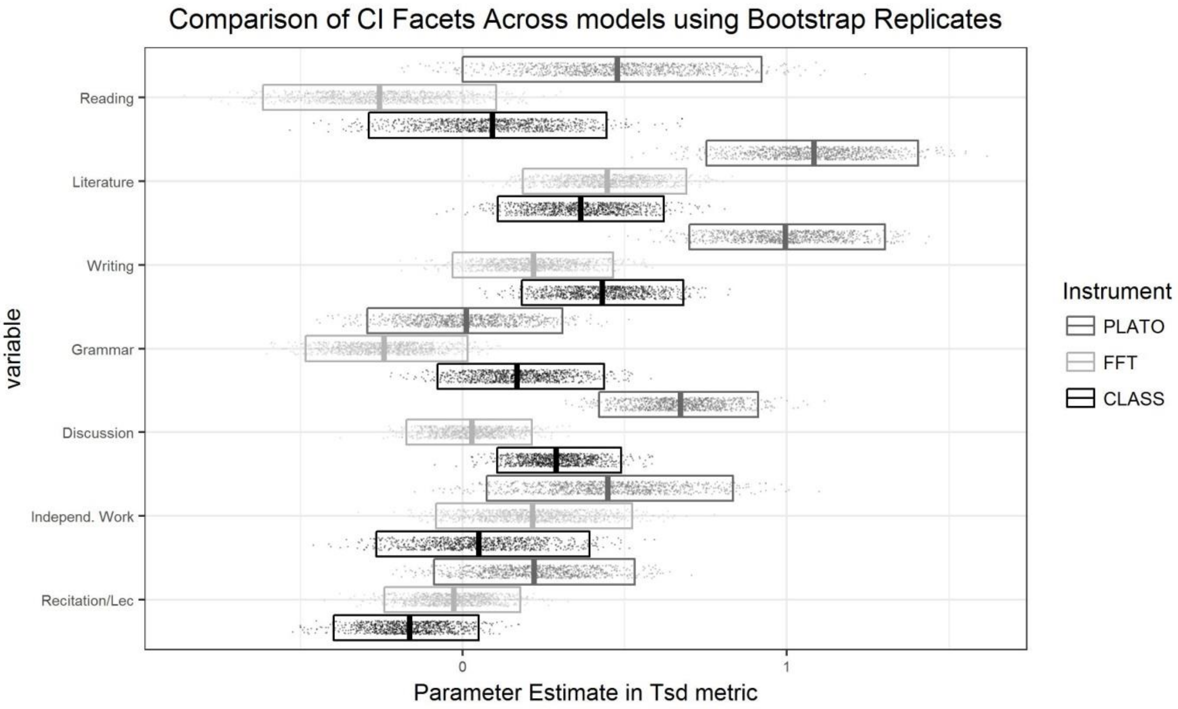

Figure F.2: Comparison of CI Facet Effects across Bootstrap Replicates

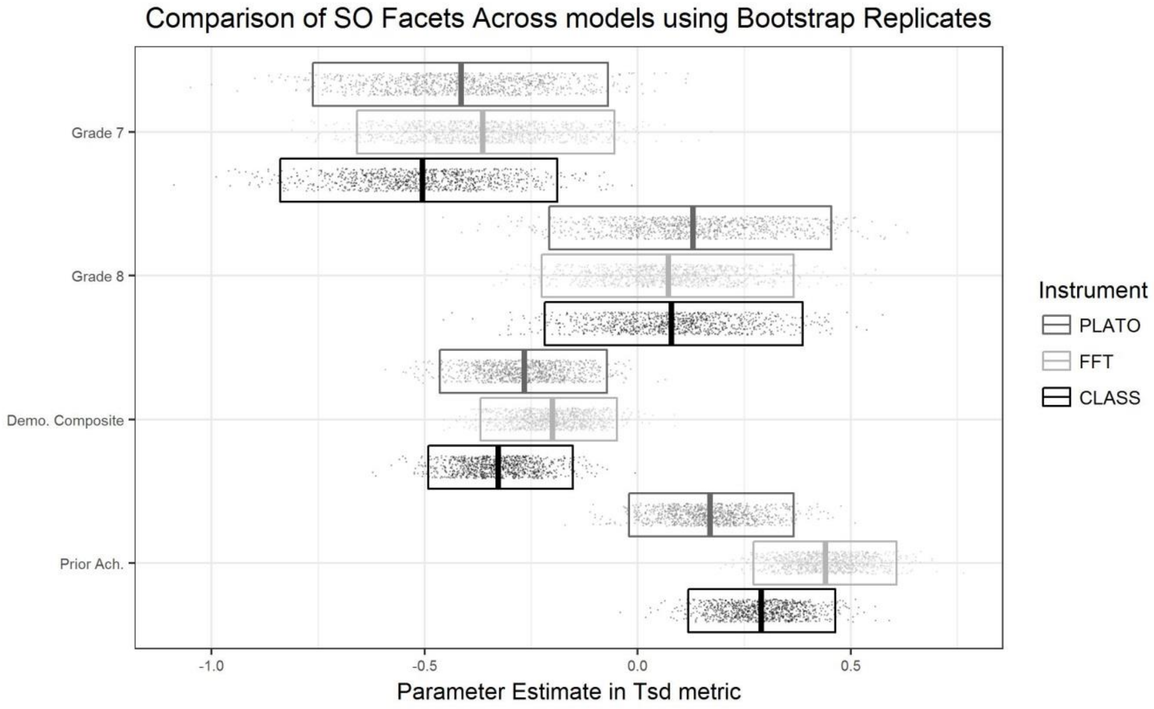

Figure F.3: Comparison of SD Facet Effects across Bootstrap Replicates

Appendix G – Comparison of Methods of Calculating Confidence Intervals for Variance Components

In order to estimate the uncertainty in the variance components of the GTheory models, I used a fully parametric bootstrap to obtain the sampling distribution of the variance components (Davison & Hinkley, 1997; Efron & Tibshirani, 1994). The parametric bootstrap assumes that the estimated model is correct and then samples from distributions of the estimated measurement facets and the residual to create a new, artificial sample. I chose a parametric bootstrap because the partially crossed nature of the data made a non-parametric bootstrap infeasible and evidence from GTheory suggests semi-parametric models tend to have biased results (Brennan, 2001). The model is used to generate estimates of observed teaching quality for this new sample and the original model is fit to this new sample. This process is repeated 1,000 times, giving 1,000 estimates of each parameter from models fit to each of the 1,000 independent, artificial bootstrapped samples. These 1,000 replicates form the sampling distribution for data equivalent in structure to the UTQ data of the GTheory model. Under the assumption that the estimated model is correct, this should estimate the sampling distribution of the population parameter of interest (e.g. the variance in observed scores that are explainable by the teacher facet: $var[\upsilon_t]$). Figure G.1 shows the distribution of these parameters across the 1,000 bootstrap replications. These distributions should be approximately normally distributed. As Figure G.1 shows, this is the case, except for when the measurement facets are estimated to have near 0 variance (i.e. near the boundary of allowable values), in which case the distributions appear approximately exponential.

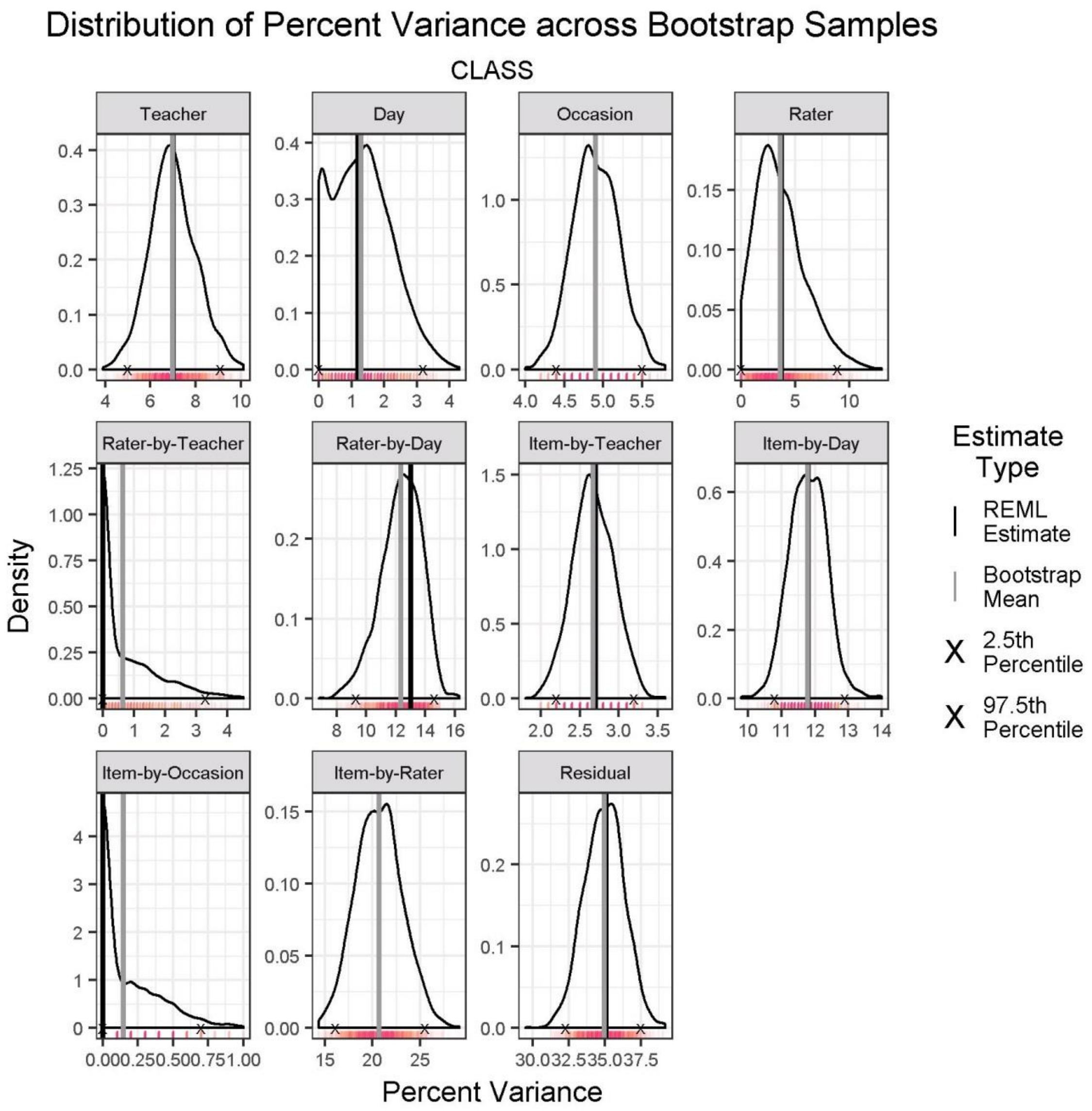

Figure G.1: Distribution of the Bootstrap Replicates for the Percentage of Variance Attributable to each Measurement Facet on CLASS

When estimating the uncertainty in parameter estimate from the bootstrap replicates, it is necessary to find a pivot (or near pivot). A pivot is a parameter whose distribution is independent of its value (Efron & Tibshirani, 1994). This is necessary so that any uncertainty in estimating the parameter of interest does not affect the estimate of the uncertainty in that parameter. That is, even if the GTheory model provides an incorrect estimate of the parameter, it must be able to provide a correct estimate of the uncertainty in that parameter for the bootstrap to work. A number of approaches exist to find pivots that can be used to convert the bootstrap replicates into the desired confidence intervals. The basic bootstrap uses the bias in the parameter estimate as a approximate pivot (i.e. $\widehat{var}[\upsilon_t] - var[\upsilon_t]$). The probability of this pivot is then used to estimate the confidence interval (i.e. $\Pr[\widehat{var}(\upsilon_t) - var(\upsilon_t)]$ is inverted substituting the original model estimate for true value and bootstrap samples for estimate). The normal bootstrap estimates a standard error of the parameter from the distribution of the bootstrap replicates and uses this to generate a confidence interval under the assumption that parameter has a normal distribution (Davison & Hinkley, 1997). This simply assumes the distribution of the parameter being estimated is normal. On the other hand, the percentile method directly uses the 95th percentile of the bootstrap replicates as an estimate of the confidence interval. It is justified in that it provides correct confidence intervals whenever a function of the parameter of interest is approximately normal (Efron & Tibshirani, 1994). All three of these bootstrap approaches are "first order accurate", which means they converge on the order of n^-0.5 (Efron & Tibshirani, 1994; Hesterberg, 2015).

Other, more complicated and more efficient approaches exist (such as the BC_a approach and studentized bootstrap), but these generally require additional knowledge of the parameter's distribution, such as the standard deviation of the distribution or an acceleration constant (Davison & Hinkley, 1997). Beyond the computationally prohibitive double bootstrap or jackknifed bootstrap, I have no way of estimating these additional values. Based on the three available approaches, the percentile bootstrap appears the best because the distributions of the parameters are not always normal or symmetric, which rules out the normal and basic bootstraps. That said, it is not clear how accurate the percentile bootstrap will be. Further, most research on the properties of these parameters is based on non-parametrically bootstrapped replicates and so may not directly apply to this problem, which uses a fully parametric approach (which has implications for how well the sampling distribution of the data is being represented).

An alternative approach to estimating confidence intervals for the estimates of the variance of measurement facets would be to use profiling (Bates, et al., 2015). Profiling uses the shape of the likelihood curve around the estimated parameters to determine how much parameters can be adjusted before a significant decrease in model fit occurs. This provides an alternative approach to exploring the uncertainty in model parameters without resorting to a bootstrap. However, I am not directly interested in the confidence intervals on the model parameter (i.e. (υ_t)), but on a function of this parameter, namely, the percentage of total variance (i.e. $var(\upsilon_t)/var(X_{\{ir(o:d:s:t)\}})$). There is no simple way to convert the profiled confidence intervals of the parameter (i.e. (υ_t)) into a percentage because the uncertainty in the variance of the observed scores is unknown and is related to the uncertainty in the individual parameters in potentially complex ways. However, I can use the profiled confidence intervals to compare the different bootstrap approaches, which I do in Figures G.2-G.4. As the figures show, there were some differences in the three bootstrap approaches and the profiled confidence intervals, but these differences are generally small. Further, the percentile bootstrap appears to be the most similar to the profiled confidence intervals, and so again is, in some sense, preferred. The percentile method is also scale invariant and so is arguably the most appropriate for the rescaled percentage of variance estimate that I am most interested in (Davison & Hinkley, 1997).

In figures G.5-G.7, I show the confidence intervals for the percentage of the total variance attributable to each measurement facet. Again, there were minimal differences across the three methods. This suggests that there is limited reason to prefer one method over the other for this problem and provides some evidence towards the robustness of results to this choice (given a number of approaches with different assumptions led to similar outcomes). I chose to use the percentile method because of its scale invariance and closer connection to the profiled confidence intervals in the raw variance estimates.

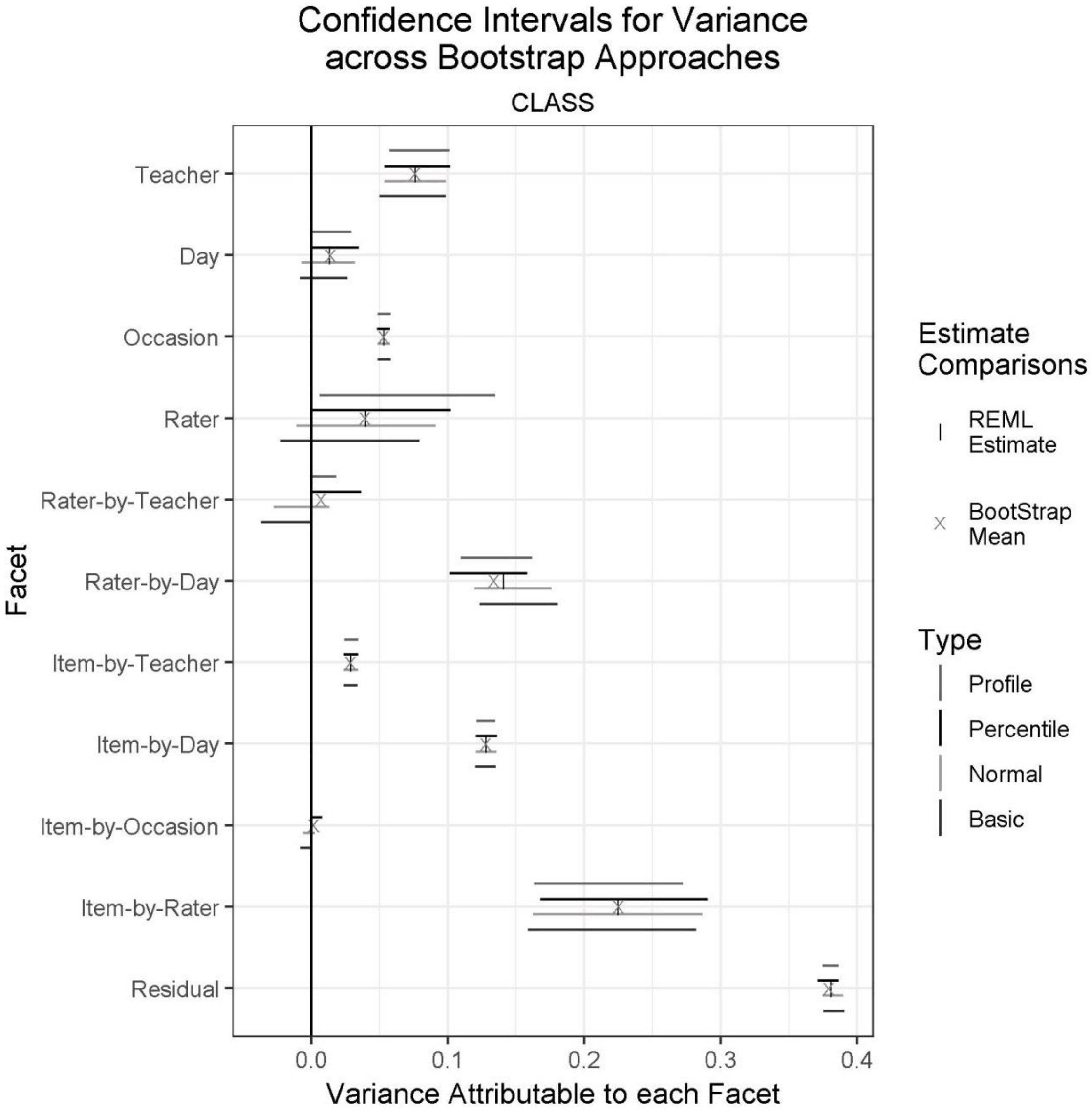

Figure G.2: Comparison of Bootstrap Confidence Intervals for the Variance of Measurement Facets on CLASS Instrument

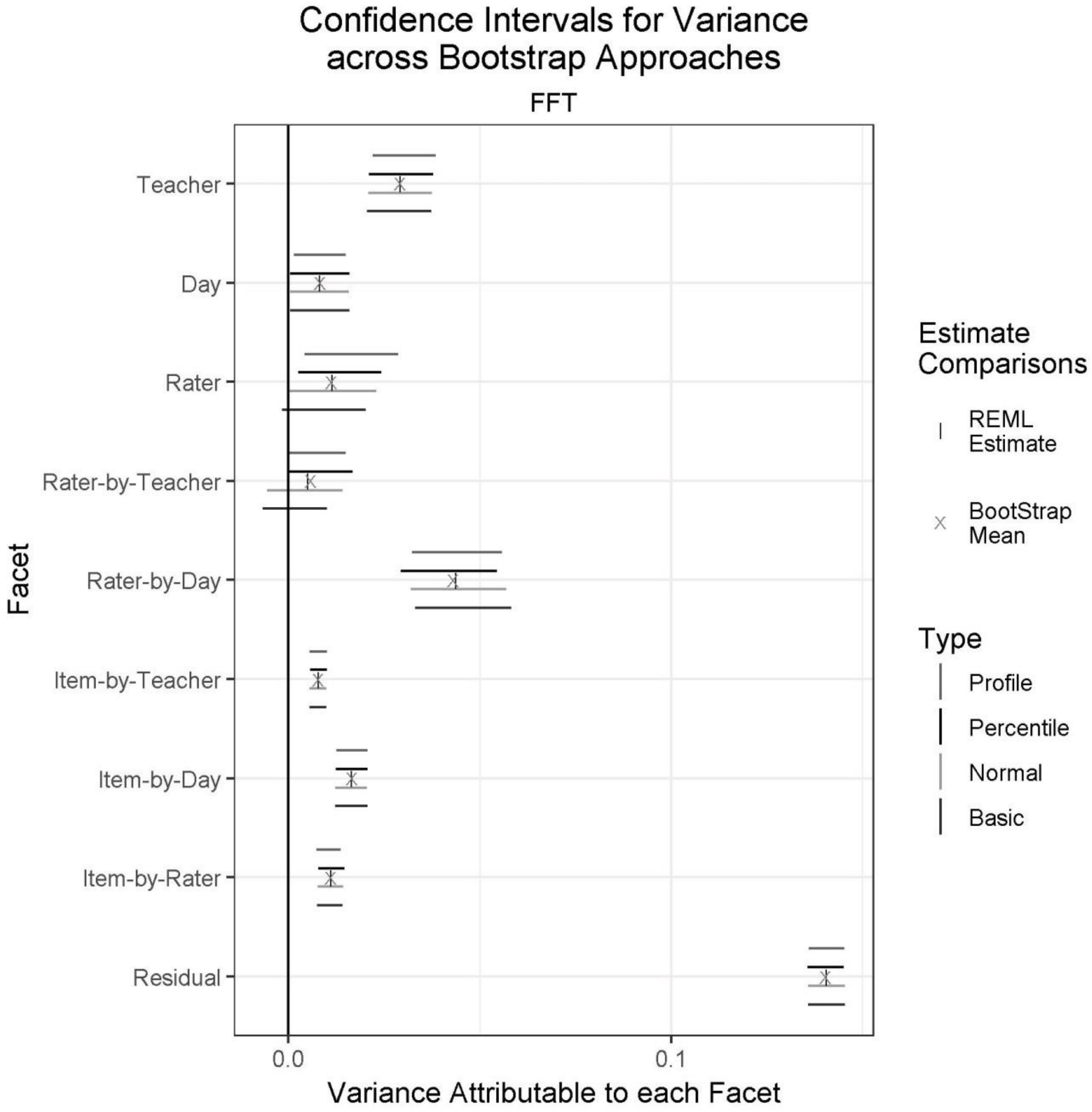

Figure G.3: Comparison of Bootstrap Confidence Intervals for the Variance of Measurement Facets on FFT Instrument

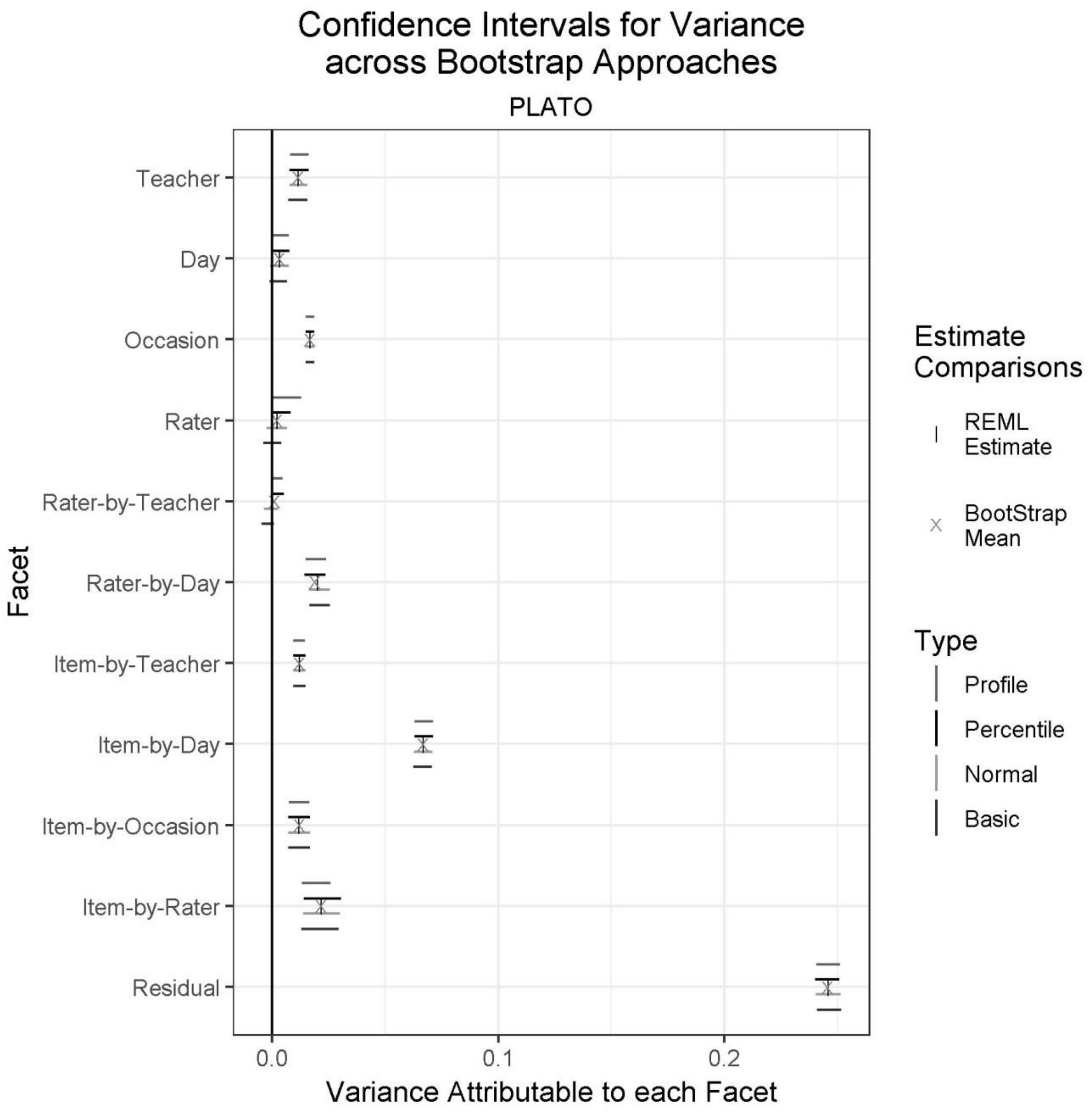

Figure G.4: Comparison of Bootstrap Confidence Intervals for the Variance of Measurement Facets on PLATO Instrument

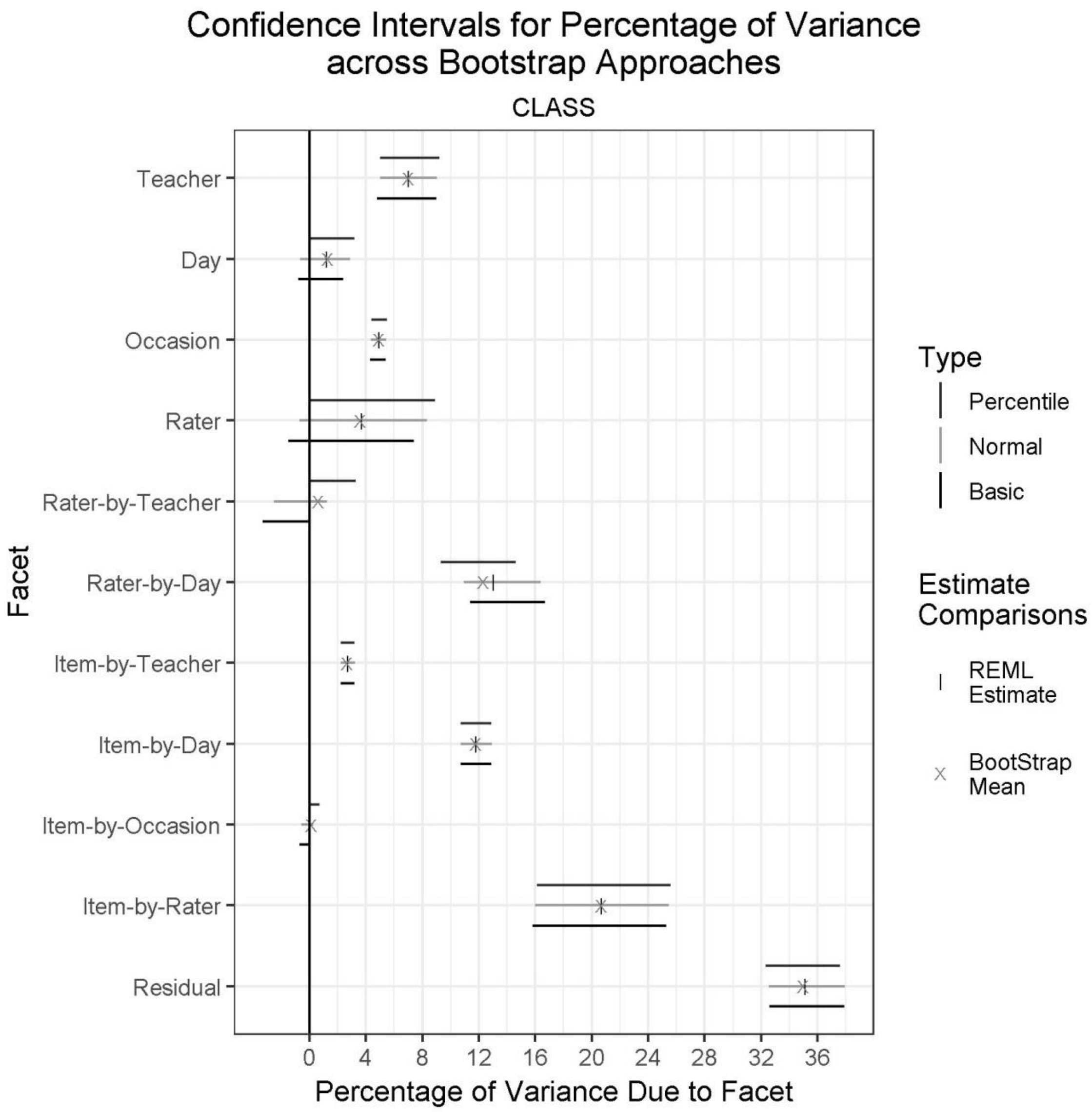

Figure G.5: Comparison of Bootstrap Confidence Intervals for the Percentage of Variance in Observed Scores due to each Measurement Facets on CLASS Instrument

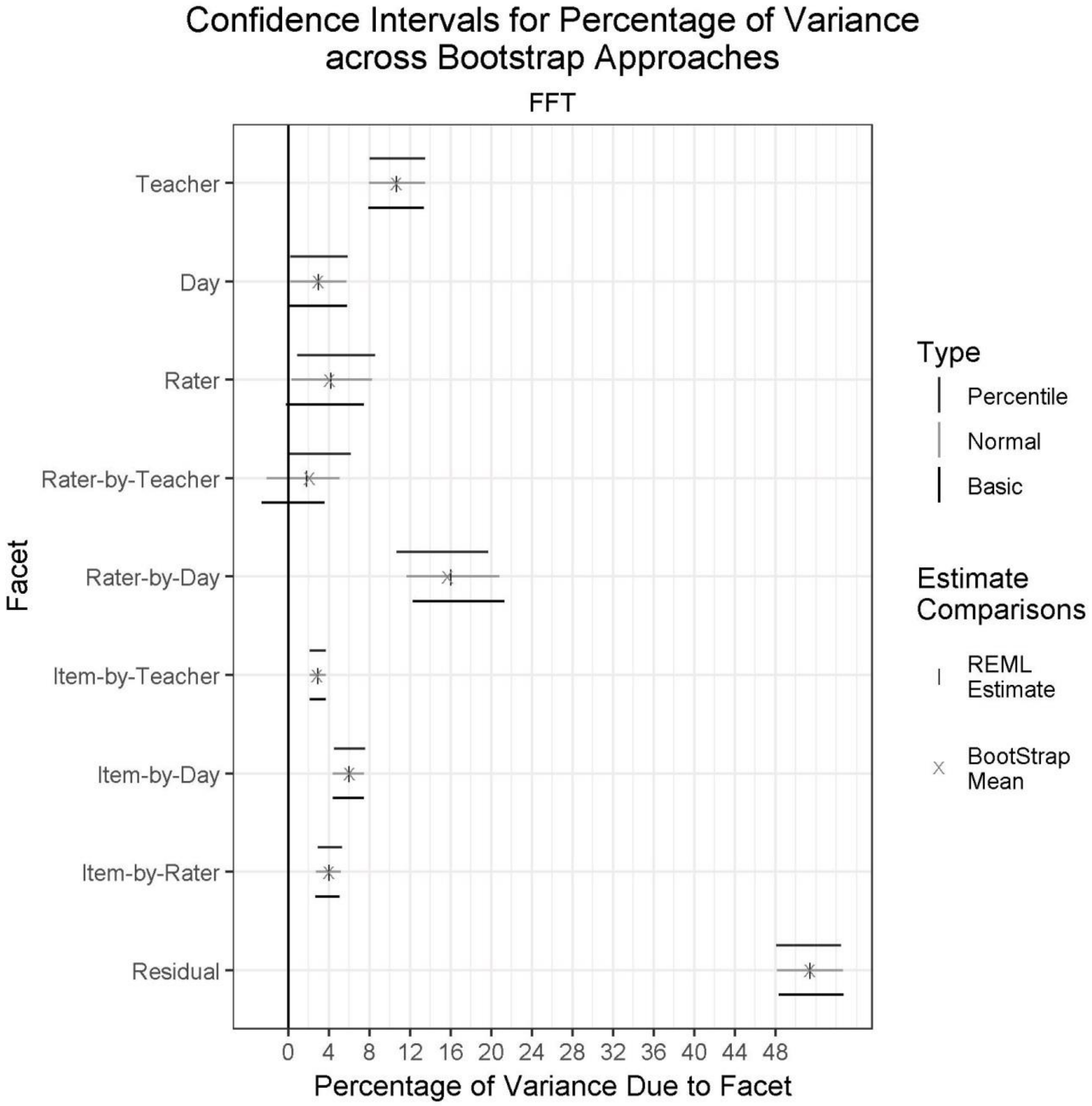

Figure G.6: Comparison of Bootstrap Confidence Intervals for the Percentage of Variance in Observed Scores due to each Measurement Facets on FFT Instrument

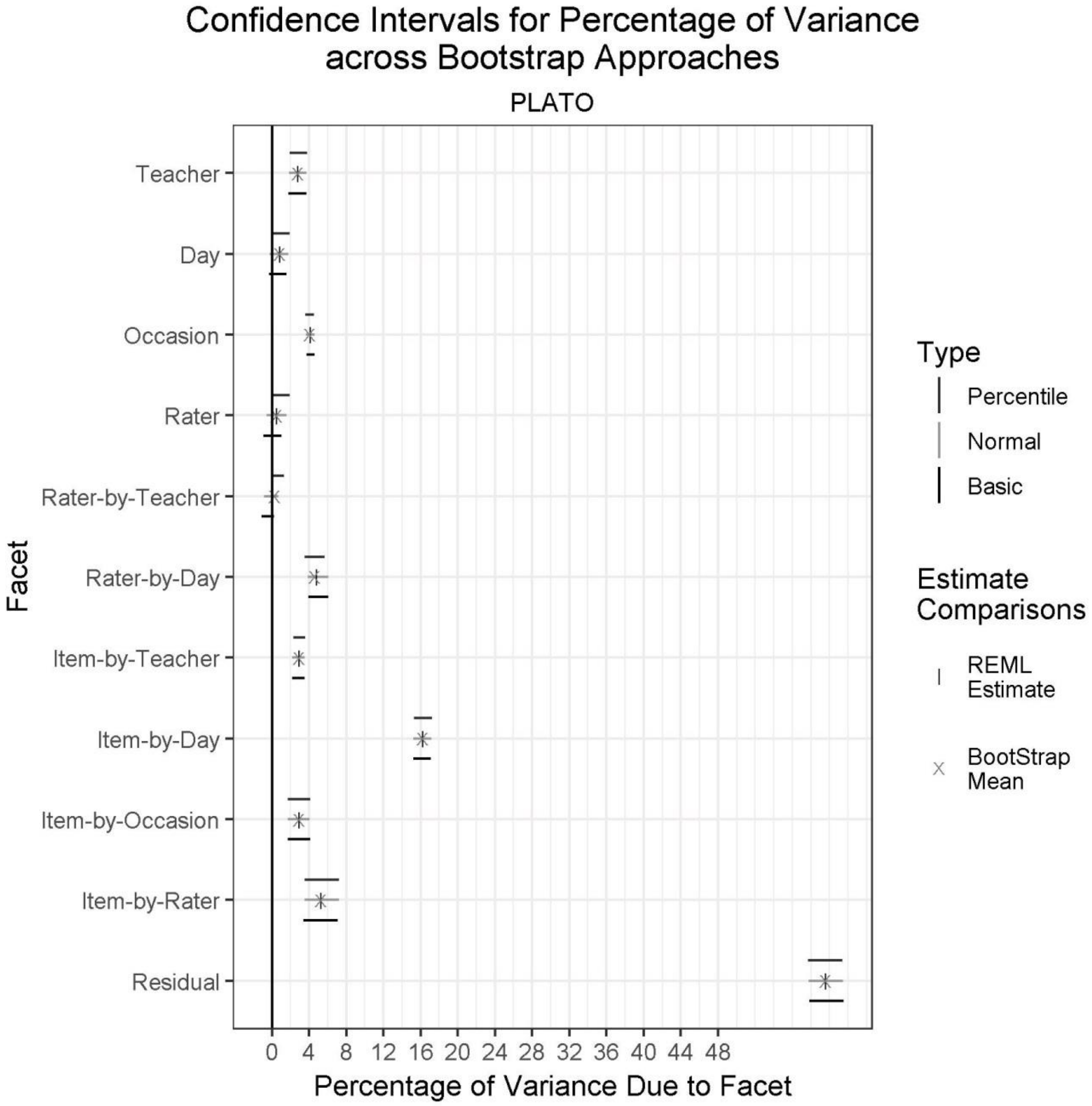

Figure G.7: Comparison of Bootstrap Confidence Intervals for the Percentage of Variance in Observed Scores due to each Measurement Facets on PLATO Instrument