USING OBSERVATION PROTOCOL SCORES TO MAKE INFERENCES ABOUT CHANGE IN TEACHER PRACTICES

by

JESSICA LYNN ALZEN

B.A., California Baptist University, 2005

B.S., California Baptist University, 2005

M.A., California Baptist University, 2007

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

School of Education

Department of Research and Evaluation Methodology

2017

This thesis entitled:
Using Observation Protocol Scores to Make Inferences about Change in Teacher Practices
written by Jessica Lynn Alzen
has been approved for the School of Education
Department of Research and Evaluation Methodology

_____

Dr. Derek Briggs

_____

Dr. Allison Atteberry

_____

Dr. Lorrie Shepard

_____

Dr. Benjamin Shear

_____

Dr. Stefanie Mollborn

Date_____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Alzen, Jessica Lynn (PhD., Research and Evaluation Methodology School of Education)

Using Observation Protocol Scores to Make Inferences about Change in Teacher Practices

Thesis directed by Professor Derek Briggs

Research on teacher learning and teacher change indicates that it is not unreasonable to expect teachers at all stages of their careers to change in their practices. However, measures of such change traditionally take the form of self-reflection and observation following teacher preparation programs (Grossman, Valencia, Evans, Thompson, Martin, & Place, 2000) or survey responses following professional development activities (Garet, Porter, Desimone, Birman, & Yoon, 2001). Even though observation protocols historically serve as the method for measuring teacher practices generally, they have yet to be used to understand change in teacher practices over time due to data limitations.

Recent changes in teacher evaluation systems initiated more frequent and consistent teacher observations (Doherty & Stevens, 2015), so change in teacher practices as measured by observation protocols might soon be of greater interest to researchers and school leaders alike. Fortunately, the Measures of Effective Teaching (MET) project provides sufficient data for beginning to understand longitudinal changes in teacher practices. The two key contributions of this project are first, an application of hierarchical linear models to estimate growth over time in teacher observation scores and second, a careful investigation of the conditions that maximize the reliability of those growth estimates.

The findings of this study suggest that teacher observation scores may change by about half of a standard deviation during a two-year time span for a few teachers, but most will show much more modest rates of growth. Further, the reliability of the growth parameter estimates can

reach as high as 0.5, but the number and spacing of observation occasions as well as number of raters required to reach such levels of reliability may be too high for practical use in some districts.

The HLM estimates in this study make an initial contribution to the research literature regarding the modeling of growth in observation scores over time. The reliability investigation provides practical information about observation system designs with the potential to yield maximally reliable estimates of growth. The former analysis gives context for future work regarding growth in observation scores while the latter informs decision-makers regarding the best choices in designing observation systems if longitudinal growth estimates are a target measure of interest.

Dedication

Whatever you do, in word or deed, do everything in the name of the Lord Jesus.

--Colossians 3:17--

Acknowledgments

Everything I have ever accomplished in this life is due to an incredible network of support. There are more people to thank than can appropriately be listed here, but I am grateful for a small space to acknowledge the sources of support most important to me.

First, my family. Not only my parents who loved me from afar these past six years, but my family with whom I live. Michael, Annabelle, Sandy, and Gary made more dinners, sacrificed more family time, and performed more than their share of household chores (ok, not Annabelle) so I could do what needed to be done to complete this project. My Michael and my Annabelle get special acknowledgement for giving me reason to find appropriate work-life balance. This dissertation could always wait for family outings and snuggles with my sweet girl. I thank them most for keeping me a well-rounded person during this process.

Second, my dearest friends who cheered me on and kept me grounded. My friends outside of academia reminded me of who I am when I am not wearing my scholar hat. I am fortunate to have many such friends, but Kim Jacobs, Wendeth Rauf, Becka Applegate and Alicia Divers deserve gold stars by their names for the number of times they reminded me I could do this. My closest fellow graduate students were also a constant lifeline, and the only ones able to give advice and support from a common experience. Ben Domingue, Nathan Dadey, Ruhan Circi, Mike Turner, Raj Chattergoon, and Amy Burkhardt, thanks for always talking me off the ledge, humoring me in the most basic of conceptual conversations, and truly being the best weapon against the imposter syndrome I constantly battle.

Third, my academic mentors. All of the members of my committee broadly, but especially Lorrie Shepard, Allison Atteberry, and Derek Briggs. Lorrie always had a way of correcting me without even slightly discouraging me, and I have always been impressed by that

quality. Allison is not only my role model for being a woman in the world of academia but additionally for just being a regular person who also happens to be an academic. This is not something one finds often in academia, at least in my experience, and knowing this about her is invaluable to me. Derek always took the time to make me better, regardless of how time-consuming or painful it was for either of us. I am better than I ever would have been had he not been my advisor. I am thankful for the way he was always considerate and supportive of my family in this whole process as well. Starting a family in graduate school was intimidating, but he never made it feel like I made a poor choice. For that, I am very grateful.

Finally, I want to acknowledge two sources of outside funding that allowed me to pay more focused attention on this project during different periods of my graduate career. First, the Miramontes Fellowship from the School of Education. Thank you to Bill and Connie Hoon Barclay for funding this fellowship. Second, the American Educational Research Association and their support of the MET Dissertation Fellowship. The funding and professional development I received from that program greatly shaped my confidence as an academic.

CONTENTS

CHAPTER

TABLES

FIGURES

1.0 Introduction

Observation protocols are evaluation tools that provide information about the practices of classroom teachers. Teacher evaluation systems use this information as indicators of teacher quality, but the implementation of observation protocols is inconsistent across the country. In 2015, 48 states required that teacher evaluations include formal observations. However, these evaluations were not necessarily annual requirements, nor were the number of observations included in evaluations standard across sates. Only 27 states required annual teacher evaluations without exception. Additionally, only 11 states required multiple annual observations as part of all teacher evaluations, while another 27 states required multiple observations for only some teacher evaluations (Doherty & Stevens, 2013; 2015). Although the regularity and frequency of teacher observations is on the rise, the lack of consistent and recurrent evaluation data from observation protocols in many states has historically made it very difficult to use observation scores as a measurement of how teachers might change or grow in their classroom practices through their careers. In other words, the data necessary for identifying the way observation scores rise or fall over time have not existed in the past.

As more states adopt multiple annual observations as part of their teacher evaluation systems, the desire to understand change in those scores throughout a year will grow. Districts investing resources in multiple annual observations may want to include information about those observations individually as well as collectively within or across years as part of multiple measures for quality in teacher and school evaluation systems. A district might also be interested in studying change in teacher observation scores throughout a year or across multiple years following programs such as district-wide professional development or new teacher mentoring systems. Opportunities for rich study of observation scores over time may be possible in the near

future as data collection continues to grow. However, there is no existing research to inform such a use of observation scores since the availability of consistent longitudinal observation scores is sparse. Fortunately, one recent large-scale study collected sufficient data from observation protocols to begin such investigations: The Measures of Effective Teaching (MET) project. The MET project, funded by the Bill and Melinda Gates Foundation, is currently the largest collection of teacher evaluation data to date. The data set is the largest in the sense that it includes a variety of measures of teacher quality from multiple districts over two years.

For the MET project, researchers set out to understand the best way to identify great teachers. In order to do this, they collected student and teacher surveys, classroom observation data, and student test scores for over 1,300 teachers across six US school districts in grades four through nine (Kane, McCaffrey, Miller, & Staiger, 2013; Kane & Staiger; 2012). The researchers collected classroom observation data eight times in each subject (math and English language arts) over a two-year time-span for up to sixteen observations per elementary teacher and eight observations per secondary teacher. The current study leverages this unique collection of data to model growth in observation scores over time.

In this dissertation, the terms "teacher practices" and "teacher behaviors" both refer to the specific practices or behaviors teachers enact in their classrooms. These are also the areas of teaching scored on observation protocols. For instance, the ways teachers manage student behavior or classroom procedures, or the ways teachers engage students in learning or use assessments to inform their instruction all exemplify teacher practices or behaviors. Districts enact a variety of activities specifically designed to help teachers improve these practices. For instance, school districts provide professional development in-services for their teachers, and content-level teams observe and critique one another's classroom practices. Further, mentor

teachers observe and coach novice teachers regarding their classroom practices. Even the mere act of a classroom observation may cause a teacher to reflect more on her practices and make changes because of that observation. Participation in any of these activities might cause teachers to change their practices in meaningful and persistent ways, and it is possible that observation scores capture at least some of these changes. The estimation and interpretation of longitudinal growth trajectories with observation scores not only provides a way for researchers to understand how observation scores rise and fall over time but is also a gap in the research literature about teacher change that is addressed in this dissertation. Thus, this dissertation aims to answer three research questions:

1) to what extent do teacher observation scores change over the course of the approximately two years of the MET project?
2) are there significant differences in the rates of growth on observation scores for novice versus experienced teachers?
3) what conditions yield the most reliable estimates of growth over time in teacher observation scores?

1.1 Conceptual Framework

This study focuses on measuring change in teacher observation scores over time with longitudinal growth trajectories. A longitudinal growth trajectory is simply a way of quantifying change over time. There are two key assumptions necessary for the fruitful estimation of a longitudinal growth trajectory for both individual teachers and groups of teachers. The first is that it is plausible to expect teachers to change (and hopefully, improve) their practices

regardless of where they are in their careers. The second is that these changes can be measured using the scores on an observation protocol applied across multiple occasions. The literature about learning to teach provides some basis for the first assumption; an examination of the second assumption is the empirical focus of this dissertation.

Teacher learning within teacher preparation programs, the first years of teaching, and in response to professional development activities are all areas of study within the larger body of research on learning to teach. Research on the efficacy of teacher preparation programs and beginning teacher practices often focus on teacher identity development and attitudes about students. In addition, this literature often explores the practices of early-career teachers and the changes that occur in those practices as teachers transition from teacher preparation programs to the first year or two of teaching (Beck, Kosnik, & Rowsell, 2007; Feiman-Nemser, 2001; Grossman, Valencia, Evans, Thompson, Martin, & Place, 2000; Martin, 2004; Peressini, Borko, Romagnano, Knuth, & Willis, 2004; Putnam & Borko, 2000; Snyder, 2012; Thompson, Windschitl, & Braaten, 2013). This work typically consists of case studies and in-depth qualitative analysis of classroom observations, interviews, teachers' reflective journals, and classroom artifacts. Although there is an abundance of qualitative work in these areas, there are few longitudinal studies that extend beyond the first year or two of classroom practices, and due to the nature of the work, sample sizes are typically 20 or fewer cases (c.f. Beck et al. 2007; Grossman et al., 2000).

A common theme among early-career research is the breakdown in the first few years of teaching practices. The first year of teaching is markedly different from the rest for teachers (Feiman-Nemser, 1983; 2001). This year is often identified as a year of "survival," "sink or swim," or a time in which teachers are more concerned with practical or procedural issues rather

than some of the more complex issues related to deep, conceptual learning that is emphasized in teacher preparation programs. Several case studies, qualitative vignettes, and ethnographic studies of early-career teachers provide evidence for this phase of teaching (Beck et al., 2007; Feiman-Nemser, 1983; 2001; Grossman et al., 2000; Martin, 2002; Snyder, 2012). However, researchers are not in agreement regarding the length of this "survival" stage of teaching. This is partly due to an acknowledgement that this phase is not universal across teachers. Primary factors that can influence how long this stage lasts are teacher disposition, teacher preparation program, and the level of support provided in the school context. Regardless of this fluidity, researchers generally identify this phase as lasting just one or two years (Feiman-Nemser, 2001; Watzke, 2007).

Following these first years, teachers generally transition away from "survival." Although they are still working to develop their craft, teachers at this stage in their careers frequently focus less on aspects of teaching such as managing student behavior and more on other aspects such as long-term planning and more careful consideration of assessment practices. Feiman-Nemser (1983; 2001) characterizes this as a shift from teachers asking themselves "Can I…" to "How do I best…". During this time, teachers might still be surprised or feel underprepared for what might happen in their classrooms. For example, they may not predict all of the misconceptions students might have about a concept and therefore may not be prepared with appropriate responses to all of those misconceptions. This process is often further complicated by within-school churn, where teachers switch grades within a school and thus need to learn new curriculum and develop additional expertise in new sets of potential student misconceptions (Atteberry, Loeb, & Wyckoff, 2017). Similar to the research on the first few years of teaching, evidence on this span

of teaching is largely based on teachers' reflective journals, observations, and interview data (Berliner, 2001; Feiman-Nemser, 2001; Watzke, 2007).

Beginning about the fifth year of teaching is when several researchers start to suggest that teachers transition from being novice teachers to teachers who are focused on improving their craft in careful and thoughtful ways (Berliner, 2001; Feiman-Nemser, 2001). By this point in their careers, teachers typically have gained a sense of self-confidence in their craft and have developed mastery of issues related to managing student behavior and organizing physical space. They now transition to the rest of their careers where they will spend the majority of their energy thinking about whether students are learning and if the instruction is appropriate for the needs of specific groups of students. There is less evidence discussing this shift in teaching primarily because the careful research on learning to teach as teachers first enter the classroom does not typically extend much beyond the first two years of a teacher's career (Beck et al. 2007; Grossman et al., 2000; Watzke, 2007).

Table 1.1 summarizes the stages of learning to teach evident in the above research. Qualitative evidence largely suggests that teachers experience a steep learning curve during their first one or two years of teaching before transitioning into a more stable time of additional growth during years three to four. Around the fifth year of teaching, teachers switch their thinking away from procedural aspects of teaching to the long-term impacts of their instruction, but there is less empirical evidence about this later transition.

Table 1.1 Stages of Learning to Teach

Description	Beginning "Survival"	Middle "Consolidation"	Final "Mastery"
Approximate years of experience	0-2	3-4	5+
Defining characteristics	• Trial and error approach to teaching and management • Feelings of uncertainty and insecurity • Focused on daily planning	• Developed reliable teaching and management strategies • Growing confidence • Long-term planning becomes a possibility	• Mechanics of teaching and management are under control • Sense of confidence and ease • Focused on long-term planning and patterns of learning with consideration of individual students' needs

After consideration of the first few years of teaching specifically, the research on teacher learning transitions to focus more on teacher learning from professional development activities. For example, content-specific professional development might focus on ways to deeply engage all students in the curriculum, how to effectively involve students at different levels of understanding, and how to anticipate student misconceptions. Studies focused on teacher learning in the context of professional development activities such as these tend to focus on teachers' attitudes and beliefs as well as changes in teacher behaviors just like the research regarding teacher preparation programs and the early years of teacher practice.

A large difference between professional development studies and those described above is that these studies tend to include much larger sample sizes (anywhere from 75 to several thousand) and primarily rely on surveys for data collection. Additionally, rather than seeking to understand change over extended periods as in the prior literature, these studies are more immediate in their timeframes. Survey data is typically collected within several weeks or a few months following the professional development activities (Franke, Carpenter, Levi, & Fennema,

2001; Garet, Porter, Desimone, Birman, & Yoon, 2001; Ingvarson, Meirers, & Beavis, 2005; Lumpe, Czerniak, Haney, & Beltyokova, 2012; Whitworth & Chiu, 2014). In these studies, teachers usually self-report regarding their opinions on the usefulness of the professional development activities, how their beliefs or attitudes about students changed, and if, and sometimes to what extent, their classroom practices change in response to the professional development experience. Although these studies are more limited in their timeframes, they still provide evidence to suggest that teachers continue to change in their thoughts, attitudes, beliefs and practices regardless of whether they are novice or veteran teachers.

The research—on teacher learning in response to professional development activities as well as the more focused research regarding teacher practices in the first few years of their careers—provides evidence to suggest that it is reasonable to expect teachers to change their practices throughout their careers. Further, this prior research also suggests that it is reasonable to expect teachers in the very first few years of teaching to change with respect to different kinds of behaviors than those who are later in their careers. This research speaks to the fact that we expect teachers at all points in their careers to see changes in their practices over time. The next section provides a brief overview of the methodologies used to explore the second assertion of this dissertation: that we can measure changes in teacher practices using scores from an observation protocol applied over multiple occasions.

1.2 Methodological Approach

Most collections of observation scores are suitable for measuring growth over time. The simplest and most straightforward approach is with a pre-post framing. Observations occur at the beginning of the year and the end of the year, and the first score subtracted from the second is a

measure of growth in teacher practices over the course of that school year. Alternatively, observations happen once in each of two adjacent school years and the second is subtracted from the first to measure growth in teacher practices across years. Although this approach provides a relatively straightforward measurement of growth, it has two important limitations. First, if there are interesting differences in the paths that teachers take in going from one score point to another, these differences are by definition unobservable (Rogosa, Brandt, & Zimowski, 1982; Willett, 1994). Second, gain scores from only two data points can suffer from low reliability (Willet, 1994; Rogosa & Willett, 1983; Cronbach & Furby, 1970).

Alternatively, longitudinal growth trajectories model growth by identifying a baseline or beginning point and then using multiple data points to detect the rate of change from that point in time over the course of the data collection period. Not only do the multiple time points provide information about what happens to a teacher's classroom practices moving from one time point to the other, but the use of multiple time points over a specified space of time (as opposed to just two) yields more reliable measures of growth (Willett, 1998). This is important, as the reliability of a measure is what allows us to use the measures to make meaningful distinctions between individuals[1] (Haertel, 2006).

Identifying salient differences in the rates of growth in teacher practices might be useful to school and district leaders for decisions such as identifying teacher leaders, assigning teachers to particular professional development activities, or using the information as part of more comprehensive teacher evaluation systems. With these goals in mind, I use a hierarchical linear

---

[1] Although this dissertation focuses on reliability, it is important to remember that precision of estimates is more useful in some contexts. This will be a topic of further discussion in the final chapter of this dissertation.

model to estimate longitudinal growth trajectories for teachers based on their observation scores in this dissertation.

1.2.1 Hierarchical Linear Modeling. A longitudinal growth trajectory is a model for growth over time. The simplest longitudinal growth trajectory is a straight line and as such, provides a parameterization of both a starting point (the intercept) and the average rate of change over time (the slope of the line). The method used in this dissertation to estimate these slopes and intercepts is hierarchical linear modeling (HLM; Raudenbush and Bryk, 2002).

HLM is a statistical method for treating hierarchical, or nested, data. Nested data means that the data cluster together into groups. For example, repeated measures of observation scores cluster within teacher. We can use HLM to understand how observation scores might change across those repeated measures (occasions) within teacher. Models that do not appropriately consider the clustered nature of the data ignore important dependencies in the error structure and violate the ordinary least squares regression assumption of independently distributed errors. This can lead to inaccurate statistical estimates if not properly addressed. Fortunately, HLM takes this clustering into account. Additionally, and perhaps more importantly in the current study, HLM allows for exploring the extent to which growth varies across teachers and for testing hypotheses regarding the factors responsible for this variability.

This study employs a two-level HLM (i.e. occasions nested within teachers) for two major purposes. The first is to explore the hypothesis that there is evidence of change in classroom practices across teachers, and to identify if novice teachers change at different rates than experienced teachers. The research literature discussed above supports the definition of novice teachers as those in their first two years of teaching and experienced teachers as those

with three or more years of experience. That same literature also suggests that novice teachers change more rapidly on practices such as those related to classroom management and discipline rather than others. Thus, the investigation of growth trajectories occurs both for overall observation scores as well as for individual dimension-level (item) scores.

The second purpose in using HLM is to explore how different observation system contexts affect the reliability of growth parameter estimates. As is later discussed in further detail, the number and spacing of observation occasions directly influences the reliability of growth trajectories. HLM allows us to explore the impact of the different combinations of number and spacing of occasions on the reliability of growth estimates. A major concern in any context is the reliability of "scores" (whether based on the aggregation of student test scores or from observation protocols) used for evaluative purposes, and the discussion regarding measures for teacher evaluation is no exception. For example, the reliability of value-added scores has been the focus of much prior research (Chetty, Friedman, & Rockoff, 2014; Kane & Staiger, 2012; McCaffrey, Sass, Lockwood, & Mihaly; Rothstein, 2010). Additionally, the reliability of observation scores on any given occasion (also referred to as status scores) is emphasized in prior research literature as well (Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013; Kane & Staiger, 2012). Since the current study is just a first step in understanding change in teacher practices over time, the best use of the results is for informative rather than evaluative purposes. Hence, a major portion of this study aimed at understanding observation system designs that yield the most reliable growth parameter estimates.

1.2.2 Reliability of growth trajectories. The reliability of growth parameters is dependent on three factors: the variance of the within-teacher (level-1) error term in the HLM,

the variance of the growth parameter estimate itself, and the number and spacing of observation occasions. The data used in the current project includes up to eight occasions per subject across two years with one observation protocol. However, the number and spacing of these occasions differ across teachers due to individual scheduling and availability constraints. As such, each teacher provides an example of a different measurement design for estimating growth over time and a respective estimate of reliability. Since this study is the first to investigate the extent to which observation scores change over time[2], there is no prior research in this area. Thus, there is no way to contextualize the findings from the current study with other findings in the field. These closing investigations of reliability not only provide a framework for better understanding the findings of the current study but also provide helpful information about how choices in implementing observation systems influence the reliability of growth measures with observation scores.

1.3 Dissertation Summary

Chapter 2 of this dissertation summarizes the relevant literature regarding teacher change generally, the use of observation protocols for teacher evaluation, and the specific observation protocol used to measure teacher change in this study. First, Chapter 2 provides a discussion of the ways teacher change beyond the first two years of teaching is typically addressed in the research literature before then providing a brief overview of the historical use of observation

---

[2] There is prior literature regarding the stability of teacher behavior (c.f. Borich, 1977; Rogosa, Floden, & Willett, 1984), but this prior work does not focus on the magnitude of change over time are evidence of teacher growth in explicit classroom practices.

scores in teacher evaluations. Next, Chapter 2 presents detailed information regarding the development of the Framework for Teaching (FFT) observation protocol as well as prior research regarding the validity and reliability of the FFT. Chapter 3 describes the data used in this project. The chapter opens with a description of the data cleaning process as well as some descriptive statistics. Next is a discussion of the process of video collection as well as specifics regarding rater recruitment and training as well as how they assigned FFT scores in the MET project. The chapter concludes with a description of the data structure used for the HLM analysis. Chapter 4 then outlines the use of individual scores to develop growth trajectories with hierarchical linear models. The analysis includes three models used to characterize growth in FFT scores over time. There is also a brief investigation into the validity of these growth estimates. The chapter concludes with a discussion of the reliability of growth parameters. Chapter 5 begins with the results from the longitudinal growth modeling of teacher practices over time. Next is a discussion of the reliability of growth estimates from a variety of contexts and related information that provides a relative framework for interpreting the findings in the current study. Recommendations are made about how local education agencies can use these findings to inform the design of their own observation systems. The dissertation closes in Chapter 6 by considering the results from a policy context as well as mentioning limitations to the current study and ideas for future work.

2.0 Literature Review

This chapter presents the prior research relevant to the current study. By the close of this literature review, the reader should have a good understanding of the prior research regarding (1) change in teacher practices; (2) the historical use of observation protocol scores; and (3) the development, validity, and reliability of the specific observation protocol used in this study.

2.1 Teacher Change

In addition to the research regarding how teachers change in the first few years of their careers discussed in the Introduction chapter of this dissertation, a healthy body of research documents the evidence regarding the ways teachers change in their practices in response to professional development experiences. This research often focuses on outcomes such as self-report regarding general change in classroom practices or changes in student-level outcomes following professional development activities. For example, Garet, Porter, Desimone, Birman, & Yoon (2001) conducted a study in which they surveyed a nationally representative sample of just over one-thousand teachers. The survey asked teachers about how various elements of professional development opportunities changed their classroom practices related to the cognitive challenge of their activities, curriculum content, instructional methods, type or mix of assessment, integration of technology, and approaches to student diversity. Survey responses ranged from 0 (no change) to 3 (significant change). The researchers averaged responses to these six areas of teacher change to create a composite scale for change in teacher practices related to professional development activities. Garet et al. found that when teachers reported gaining

enhanced knowledge and skills from professional development opportunities, they also reported positive changes in their teaching practices (0.44 unit change, on average). In addition, when teachers reported alignment between particular professional development activities and other professional development experiences, standards, and assessments, they were even more likely to change their practices (an additional 0.21 unit change, on average). Although these are positive results, since the measures ranged from no change in practices to significant change in practices it is unclear *how* teacher behaviors changed and if that change was necessarily for the better. In addition, the measures are all based on self-report. These findings convey information about the efficacy of the professional development in causing teachers to believe they changed their practices as opposed to providing insight into the type, quality, and extent of change in specific teacher practices.

In another study, Lumpe, Czerniak, Haney, & Beltyokova (2012) investigated how teacher beliefs improved after participating in professional development and if those beliefs were predictive of student achievement. Lumpe et al.'s study included approximately 450 elementary school teachers, each of whom participated in a two-week long summer professional development seminar (a total of 80 contact hours) focused on inquiry-based instruction, science content knowledge, and science process. An additional twenty-four hours of professional development activities occurred over the course of the academic year through activities such as peer coaching and meeting with support teachers. Following the first year of the study, teachers completed a survey designed to capture their beliefs about teaching science. In addition, the researchers used data from state standardized tests in science as a measure of student achievement. Findings in this study included teachers having significantly more positive self-efficacy beliefs about teaching science ($t=12.03$, $p<.001$) following participation in professional

development activities. Further analysis indicated that both teacher self-efficacy beliefs and the number of hours of professional development were significant predictors of fourth-grade student achievement outcomes.

In a different approach, Harootunian & Yargar, (1980) surveyed about 240 K-12 teachers in a single district with up to 39 years of experience. Teachers listed events or changes in their teaching that indicated success to them. About three out of every four responses recorded by the teachers defined success with relation to something students did (e.g. students listening, enthusiasm of students, or good grades on quizzes and tests) as opposed to something that they themselves did (e.g. enjoying teaching a difficult concept, showing firmness and fairness with students). Similarly, Lavigne & Bozack (2015) surveyed seventy-five teachers in grades K-9 in a large urban school district in the Midwest. They found that the proportion of self-focused to student-focused responses followed a similar pattern to the earlier Harootunian & Yargar (1980) study. That is, teachers identified successes related to changes in student behavior much more frequently than related to changes in their own behavior.

Previous research often frames teacher change as evidence of the efficacy of professional development activities as opposed to a characteristic of teachers themselves. The recent upswing in the usage of observation protocols not only provides longitudinal data that were previously unavailable, but also enables an extension of the research on teacher change to include study of systematic changes in specific teacher practices as measured by observation protocols. Teacher evaluations historically include observation scores as the single or one of several measures of teacher quality. The current project suggests a new way of using observation scores to gain information about teacher quality. Applying longitudinal growth trajectories to observation scores provides information about the ways and extent to which teachers change in their

practices. This is valuable information both for gaining a broader perspective on teacher quality in general but also for uses such as identifying the most appropriate professional development activities or strong teacher leaders.

2.2 Historical Use of Observation Protocols

Observation protocols are historically the most pervasive and persistent measure for teacher evaluations (Hill, Charalambous, & Kraft, 2012). However, there are two common critiques of the practice: 1) in most historical circumstances, teachers received reviews based on a single classroom observation conducted by an administrator with little or no training in how to adequately perform observations, and 2) there is typically little variability in observation scores, with upwards of 98% of teachers receiving ratings of satisfactory or higher in any given context (Braun, 2005; Milanowski, 2011; Kane & Staiger, 2012; Weisberg, Sexton, Mulhern, & Keeling, 2009). For example, Weisberg et al. (2009) conducted a survey of approximately 15,000 teachers across 12 districts in four states. Of these teachers, 99% of them received observation scores of satisfactory or higher. The crux of the second critique is that other measures of teacher quality, such as value-added scores, indicate more variability in teacher quality (c.f. Aaronson, Barrow, & Sander, 2007; Chetty, Friedman, & Rockoff, 2014; Gordon, Kane, & Staiger, 2006). A practice that consistently provides nearly all teachers with the same final ratings when other evidence suggests otherwise calls for further consideration.

In response to these criticisms, much work has been done in recent years to improve the use of observation scores for teacher evaluations. These changes were largely motivated by the 2009 Race to the Top (RttT) grant competition sponsored by the U.S. Department of Education.

This competition encouraged states to adopt rigorous teacher evaluation systems that included multiple measures of teacher effectiveness by making such an adoption a requirement for receiving the most points in the competition (U.S. Department of Education, 2009). Most evaluation systems now include multiple measures of teacher quality. In addition, other efforts focus on improving the *implementation* of observations.

The 2015 National Council on Teacher Quality report indicates that forty-eight states now require teacher evaluations to include formal observations and twenty-seven states require multiple observations for at least some teachers (Doherty & Jacobs, 2015). In addition, states and districts are devoting more resources to improve observation systems. These systems include not only the observation protocols themselves but also systems for training and certifying raters to use those protocols. Further, more consideration is paid to the influence of the number and length of observations as well as the number of raters per observation (Hill et al., 2012; Ho & Kane, 2013; Mashburn, Meyer, Allen, & Pianta, 2014). In many contexts, the Danielson Framework for Teaching is at least one, if not the primary, observation protocol that is adopted in these teacher observation systems.

2.3 The Danielson Framework for Teaching

As of 2013, over twenty states adopted the Danielson Framework for Teaching (FFT) as either the single model or one of several approved models for teacher observations (Danielson Group, 2013). The Measures of Effective Teaching (MET) Project included four different observation protocols applied to each video in the data, and the FFT was one in addition to the Classroom Assessment Scoring System (CLASS), the Protocol for Language Arts Teacher

Observations (PLATO), and the Mathematical Quality of Instruction (MQI). The CLASS and FFT are general observation protocols designed for any classroom, regardless of content area. The MQI and PLATO are subject-specific protocols designed for mathematics and language arts classrooms respectively.

The FFT is the data source used in this dissertation for multiple reasons. First, it is desirable to use a general protocol to maximize the number of potential data points for any given teacher. Over the course of the study, teachers provided up to eight videos for each subject taught. Elementary teachers, who instructed the same group of students in multiple subjects, provided up to sixteen videos over the course of the two-year study: a maximum of eight in English Language Arts (ELA) and a maximum of eight in math. Secondary teachers provided a maximum of eight videos in one subject or the other. Since scores on the FFT are agnostic as to course content, all scores, regardless of course content, are available for building longitudinal growth trajectories. Second, the FFT is preferable over the CLASS primarily because there is one version of the FFT used across grades 4-9 while the CLASS has two forms, one for grades 4-5 and another for grades 6-9. Using the FFT allowed for inclusion of all grades in the study without concern regarding change in the protocol. Finally, the FFT is one of the most extensively used general observation protocols. Since the protocol is so widely implemented, study of growth with scores from the FFT is relevant to a large audience.

2.3.1 Protocol development. Charlotte Danielson originally designed the FFT in 1996 as an extension of her work with the Educational Testing Service to develop a measure for evaluating teacher licensure applicants. From this initial project, Danielson examined current research and extended the protocol to include teaching skills expected not only for novice

teachers but for more experienced teachers as well. The first iteration as well as every subsequent iteration divided teaching into four "domains," with twenty-two "dimensions[3]" spread across these four domains. Table 2.1 provides information about the four domains, labeled as Planning & Preparation, Classroom Environment, Instruction, and Professional Responsibilities as well as the dimensions related to each domain.

When the Bill and Melinda Gates Foundation chose the FFT as one of two general observation protocols for the MET project, the Danielson Group[4] revised the protocol and provided additional training tools to help with rater training and scoring. The changes made for the MET project were largely for the purpose of clarity and ease of scoring (Danielson, 2011). For example, the MET project excluded "Organizing Physical Space" from Domain 2 and "Demonstrating Flexibility and Responsiveness" from Domain 3. The excluded dimensions required more information regarding teacher preparation outside of the immediate lesson or more information about the organization of classroom space than was available from video data. In addition, the Danielson Group developed examples for each performance level for each dimension as well as "critical attributes" to give guidance in distinguishing between levels of performance.

---

[3] The Danielson (2011) documentation refers to the dimensions as "components", but the MET documentation uses the term dimensions. I use dimensions throughout this dissertation for consistency.

[4] The consulting organization founded and run by Charlotte Danielson that manages work related to the FFT.

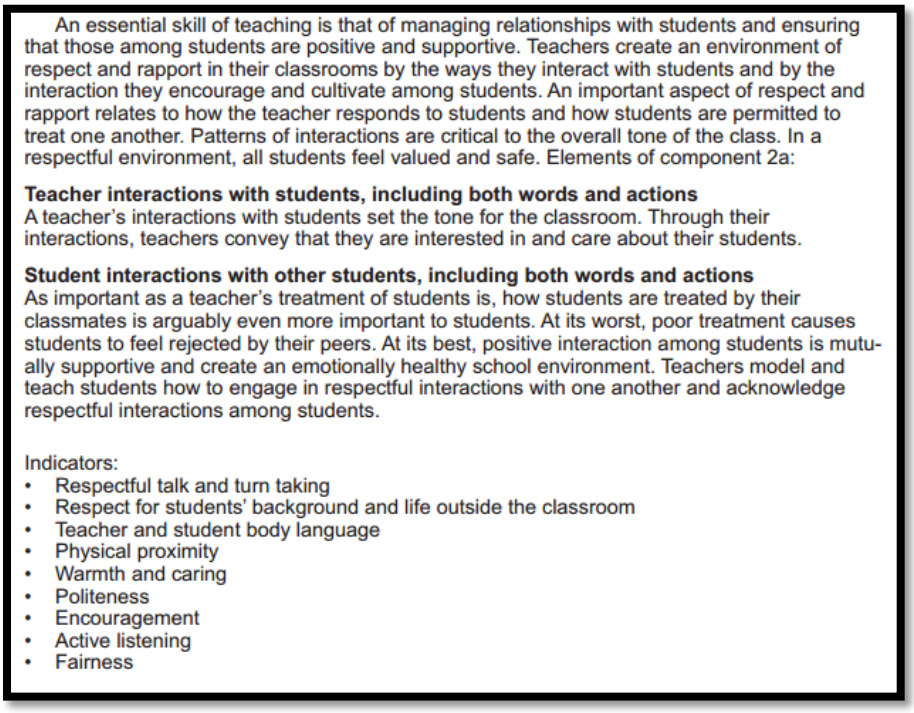
Table 2.1 FFT Domains and Dimensions[5]

Domains	Dimensions
1. Planning & Preparation	• Knowledge of Content and Pedagogy • Demonstrating Knowledge of Students • Setting Instructional Outcomes • Demonstrating Knowledge of Resources • Designing Coherent Instruction • Designing Student Assessments
2. Classroom Environment	• Creating an Environment of Respect and Rapport • Establishing a Culture for Learning • Managing Classroom Procedures • Managing Student Behavior • Organizing Physical Space
3. Instruction	• Communicating with Students • Using Questioning and Discussion Techniques • Engaging Students in Learning • Using Assessment in Instruction • Demonstrating Flexibility and Responsiveness
4. Professional Responsibilities	• Reflecting on Teaching • Maintaining Accurate Records • Communicating with Families • Participating in a Professional Community • Growing and Developing Professionally • Showing Professionalism

The FFT documentation and training materials include a full description, critical attributes, and examples for every dimension on the protocol. To illustrate the items for one of the dimensions used in the MET project, consider Figures 2.1-2.3. Figure 2.1 shows the full description from the FFT for Domain 2: Dimension 1"Creating an environment of respect and rapport." The full description provides a general orientation to the dimension as well as a

[5]In 2007, Danielson revised the FFT based on research literature from the previous decade. Some small changes were made to the component language in order to reflect more recent research. For example, what was previously called "Providing feedback to students" is now called "Using assessment in instruction" and what was previously called "Communicating clearly and accurately" is now called "Communicating with students". Revisions such as these were mostly minor changes in language for the purpose of clarifying the components of the protocol. However, the 2007 version also added elements under each dimension designed to help further define each dimension.

detailed and careful description so that raters have a better understanding of what they are attempting to capture when rating a teacher on this particular dimension. An expanded definition of the elements (in bold) nested under each dimension is provided as well. Though these elements were not assigned separate scores in the MET project specifically, they were still key factors that raters were trained to look for in classroom videos. Each description also includes a list of indicators regarding specific behaviors raters could look for in the classroom as evidence of this particular dimension.



An essential skill of teaching is that of managing relationships with students and ensuring that those among students are positive and supportive. Teachers create an environment of respect and rapport in their classrooms by the ways they interact with students and by the interaction they encourage and cultivate among students. An important aspect of respect and rapport relates to how the teacher responds to students and how students are permitted to treat one another. Patterns of interactions are critical to the overall tone of the class. In a respectful environment, all students feel valued and safe. Elements of component 2a:

Teacher interactions with students, including both words and actions
A teacher's interactions with students set the tone for the classroom. Through their interactions, teachers convey that they are interested in and care about their students.

Student interactions with other students, including both words and actions
As important as a teacher's treatment of students is, how students are treated by their classmates is arguably even more important to students. At its worst, poor treatment causes students to feel rejected by their peers. At its best, positive interaction among students is mutually supportive and create an emotionally healthy school environment. Teachers model and teach students how to engage in respectful interactions with one another and acknowledge respectful interactions among students.

Indicators:
- Respectful talk and turn taking
- Respect for students' background and life outside the classroom
- Teacher and student body language
- Physical proximity
- Warmth and caring
- Politeness
- Encouragement
- Active listening
- Fairness

Figure 2.1 Full description of "Creating an environment of respect and rapport." Adapted from Danielson Framework for Teaching (p. 28), by C. Danielson, 2011.

While Figure 2.1 is a general description of "Creating an environment of respect and rapport," Figure 2.2 subdivides parallel information about the dimension for each of the four potential scoring levels on the FFT. The four scoring levels are unsatisfactory (score of 1), basic

(score of 2), proficient (score of 3) and distinguished (score of 4). In Figure 2.2, the box immediately underneath each scoring level defines how the dimension might look at that scoring level in a broad sense. A second box underneath each scoring level includes explanations of critical attributes. These are specific examples of what exactly a rater might see in a teacher's classroom to merit the respective scoring level.

UNSATISFACTORY	**BASIC**
Patterns of classroom interactions, both between the teacher and students and among students, are mostly negative, inappropriate, or insensitive to students' ages, cultural backgrounds, and developmental levels. Interactions are characterized by sarcasm, put-downs, or conflict. Teacher does not deal with disrespectful behavior.	Patterns of classroom interactions, both between the teacher and students and among students, are generally appropriate but may reflect occasional inconsistencies, favoritism, and disregard for students' ages, cultures, and developmental levels. Students rarely demonstrate disrespect for one another. Teacher attempts to respond to disrespectful behavior, with uneven results. The net results of the interactions is neutral, conveying neither warmth nor conflict.
Critical Attributes	
Teacher uses disrespectful talk towards students; student's body language indicates feelings of hurt or insecurity. Students use disrespectful talk towards one another with no response from the teacher. Teacher displays no familiarity with or caring about individual students' interests or personalities.	The quality of interactions between teacher and students, or among students, is uneven, with occasional disrespect. Teacher attempts to respond to disrespectful behavior among students, with uneven results. Teacher attempts to make connections with individual students, but student reactions indicate that the efforts are not completely successful or are unusual.

PROFICIENT	DISTINGUISHED
Teacher-student interactions are friendly and demonstrate general caring and respect. Such interactions are appropriate to the ages of the students. Students exhibit respect for the teacher. Interactions among students are generally polite and respectful. Teacher responds successfully to disrespectful behavior among students. The net result of the interactions is polite and respectful, but impersonal.	Classroom interactions among the teacher and individual students are highly respectful, reflecting genuine warmth and caring and sensitivity to students as individuals. Students exhibit respect for the teacher and contribute to high levels of civil interaction between all members of the class. The net results of interactions is that of connections with students as individuals.
Critical Attributes	
Talk between teacher and students and among students is uniformly respectful. Teacher responds to disrespectful behavior among students. Teacher makes superficial connections with individual students.	In addition to the characteristics of "proficient": Teacher demonstrates knowledge and caring about individual students' lives beyond school. When necessary, students correct one another in their conduct toward classmates. There is no disrespectful behavior among students. The teacher's response to a student's incorrect response respects the student's dignity.

Figure 2.2 Critical attributes of "Creating an environment of respect and rapport." Adapted from Danielson Framework for Teaching (p. 30-31), by C. Danielson, 2011.

Although Figure 2.2 is the general rubric used for scoring Domain 2: Dimension 1, the Danielson framework includes one additional resource to assist raters. Figure 2.3 depicts the matrix of even more specific examples for each scoring level of the same dimension. These are examples of additional factors by which raters can justify a score at any level for Domain 2: Dimension 1.

The FFT materials include a full description, elements and indicators (Figure 2.1), score-level descriptors and critical attributes (Figure 2.2), and score-level examples (Figure 2.3) for every dimension on the protocol. The Danielson Group created these materials for the MET project, but also further developed the protocol to include these materials for the domains and dimensions not included in the MET project to help with future trainings in situations where raters would have access to more than just classroom videos.

Unsatisfactory	Basic	Proficient	Distinguished
A student slumps in his/her chair following a comment by the teacher. Students roll their eyes at a classmate's idea; the teacher does not respond. Many students talk when the teacher and other students are talking; the teacher does not correct them. Some students refuse to work with other students. Teacher does not call students by their names.	Students attend passively to the teacher, but tend to talk, pass notes, etc. when other students are talking. A few students do not engage with others in the classroom, even when put together in small groups. Students applaud halfheartedly following a classmate's presentation to the class. Teacher says: "Don't talk that way to your classmates," but student shrugs his/her shoulders.	Teacher greets students by name as they enter the class or during the lesson. The teacher gets on the same level with students, kneeling, for example, beside a student working at a desk. Students attend fully to what the teacher is saying. Students wait for classmates to finish speaking before beginning to talk. Students applaud politely following a classmate's presentation to the class. Students help each other and accept help from each other. Teacher and students use courtesies such as "please," "thank you," "excuse me." Teacher says: "Don't talk that way to your classmates," and the insults stop.	Teacher inquires about a student's soccer game last weekend (or extracurricular activities or hobbies). Students hush classmates causing a distraction while the teacher or another student is speaking. Students clap enthusiastically after one another's presentations for a job well done. The teacher says: "That's an interesting idea, Josh, but you're forgetting ..."

Figure 2.3 Examples of "Creating an environment of respect and rapport." Adapted from Danielson Framework for Teaching (p. 29), by C. Danielson, 2011.

Once certified to score videos with the FFT in the MET project (details regarding rater training and certification are discussed in the Data chapter), raters used materials like those in Figures 2.1 – 2.3 for each dimension to assign each video eight different scores (see Table 2.2). Score assignment occurred only at the dimension level, so raters did not aggregate scores to the domain level or give any form of total score for the video. The longitudinal growth analysis in

this project uses both these individual dimension-level scores as well as the average across all eight dimensions for any given occasion as a composite measure of each lesson overall.

Table 2.2 FFT Domains and Dimensions Used in the MET Project

Domain	Dimensions
2. Classroom Environment	• Creating an Environment of Respect and Rapport • Establishing a Culture for Learning • Managing Classroom Procedures • Managing Student Behavior
3. Instruction	• Communicating with Students • Using Questioning and Discussion Techniques • Engaging Students in Learning • Using Assessment in Instruction

2.3.2 Validity of FFT scores. The Standards for Education and Psychological Testing (2014) define validity of the use of scores from a test instrument as follows:

> the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself. (p. 11)

There are meaningful applications of the same standard of validity to observation protocols even though these standards specifically reference test instruments. In an observation protocol context, the dimension scores serve as the "items" on the "test" that is the protocol. As such, an argument for the valid use of scores from the FFT should meet the requirements outlined in the standards for education and psychological testing.

Prior research on various fronts exists in support of the validity of FFT scores. This section provides a brief summary of that related research. As the standards suggest is necessary for a validity argument, this section begins by providing "an explicit statement of the proposed interpretation of [observation] scores, along with a rationale for the relevance of the

interpretation to the proposed use" (AERA, APA, NCME, 2014, p. 11). The Framework for Teaching Evaluation Instrument, states that scores from the FFT

> [identify] those aspects of a teacher's responsibilities that have been documented through empirical studies and theoretical research as promoting improved student learning. Although not the only possible description of practice, these responsibilities seek to define what teachers should know and be able to do in the exercise of their profession. (Danielson, 2011, p. iv)

Danielson suggests that one interpretation of scores from the FFT is that they represent the aspects of a teacher's responsibilities related to improved student learning. She argues that though the practices captured by FFT scores are not comprehensive, those represented on this instrument indicate key behaviors that teachers should know and be able to do in performing their job responsibilities. This statement about the proposed interpretation of observation scores makes logical sense given the frequent use of FFT scores as part of multiple measures of teacher quality for teacher evaluation.

Acceptable sources of validity evidence can come from evidence local to the current context as well as the use of the instrument in other settings (AERA, APA, NCME, 2014). The sections that follow include evidence from within as well as outside of the MET project context and include evidence based on the following sources: content, internal structure, and relations to other variables.

2.3.2.1 Evidence based on content. The Standards for Education and Psychological Testing (2014) state that

> important validity evidence can be obtained from an analysis of the relationship between the content of a test and the construct it is intended to measure. […] Evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores. (p. 14)

The FFT developers claim that the teacher practices characterized by the dimensions within its four domains (Table 2.1) are salient to the teaching profession. One way to validate these dimensions on a content level is to consider alignment between these dimensions and some set of professional teaching standards. Norman Webb (1999) published an article regarding four key components of alignment between content standards and standardized tests. Although the current context is not the same as Webb's original work, there is also viable application to the context of teaching standards and observation protocols to some extent. Webb's four alignment criteria are categorical concurrence, depth of knowledge consistency, range of knowledge correspondence, and balance of representation. A full alignment study includes consideration of all four criteria, but this section focuses on categorical concurrence, as it is the most pertinent to the observation protocol setting.

Webb (1999) defines categorical concurrence as both the standards and the assessment having the same or consistent content. In practice, he specifies that a sufficient number of the categories within the standards should be evidence on the assessment tool. Correspondence with the Danielson Group about validity evidence as well as a Google search for official content alignment studies between the FFT and teaching standards—particularly for states that adopted the FFT as one of their protocols or the states or districts included in the MET study yielded no results. However, the New Jersey and New York City Departments of Education as well as the National Board of Professional Teaching Standards resource center provide basic materials that indicate the overlap between their professional teaching standards and the FFT. Washington identifies the FFT as one of the approved state observation protocols. New York City is one of the districts from the MET study and uses the protocol in their formal teacher observations. Finally, the National Board of Professional Teaching is a non-profit organization started in 1983

that does work to promote national teaching standards as well as a national teaching certificate. Their research center identified the overlap between the FFT and the national standards.

Table 2.3 provides a brief summary of the overlap between the FFT and these teaching standards. Although a complete evaluation of the alignment between these standards and the Danielson framework requires more careful reading of not only the description of the FFT dimensions but also the individual standards, Table 2.3 provides a broad overview of the content connections between the FFT dimensions and professional teaching standards in multiple contexts.

Table 2.3 Alignment of FFT Dimensions to state-Level Professional Teaching Standards

FFT Dimension	Washington[6]	New York City[7]	National Board[8]
Creating an Environment of Respect and Rapport	Criterion 5: Fostering and managing a safe, positive learning environment	Standard 1.4: Maintain a culture of mutual trust and positive attitudes that supports the academic and personal growth of students and adults	Core Proposition 1: Teachers are committed to students and their learning Core Proposition 3: Teachers are responsible for managing and monitoring students' learning. Core Proposition 4: Teachers think systematically about their practice and learn from experience
Establishing a Culture for Learning	Criterion 1: Centering instruction on high expectations for student achievement	Standard 3.4: Establish a culture for learning that communicates high expectations to staff, students, and families, and provide supports to achieve those expectations	Core Proposition 1 Core Proposition 2: Teachers know the subjects they teach and how to teach those subjects to students Core Proposition 3 Core Proposition 4

[6] Adapted from the Washington Teacher Evaluation Criteria: http://www.k12.wa.us/TPEP/Frameworks/Danielson/Danielson_WA_Alignment.pdf

[7] Adapted from the New York City "Alignment across the NYCDOE": http://schools.nyc.gov/NR/rdonlyres/7D5834A8-A01D-4D99-9F28-6E0D613EBC69/0/FrameworkforGreatSchoolsAlignmentAcrosstheNYCDOE.pdf

[8] Adapted from the National Board's Five Core Propositions: http://nbrc.illinoisstate.edu/downloads/nbrc/crosswalk/1charlottedanielson.pdf

Managing Classroom Procedures	Criterion 5	Standard 3.4	Core Proposition 1 Core Proposition 3 Core Proposition 4 Core Proposition 5: Teachers are members of learning communities
Managing Student Behavior	Criterion 2: Demonstrating effective teaching practices Criterion 5	Standard 1.4	Core Proposition 1 Core Proposition 3 Core Proposition 4
Communicating with Students	Criterion 1	Standard 1.1: Ensure engaging, rigorous, and coherent curricula in all subjects, accessible for a variety of learners.	Core Proposition 1 Core Proposition 2 Core Proposition 3 Core Proposition 4
Using Questioning and Discussion Techniques	Criterion 2	Standard 1.1	Core Proposition 1 Core Proposition 2 Core Proposition 3 Core Proposition 4
Engaging Students in Learning	Criterion 1	Standard 1.2: Develop teacher pedagogy from a coherent set of beliefs about how students learn best that is informed by the instructional shifts and Danielson Framework for Teaching, aligned to the curricula, engaging, and meets the needs of all learners so that all students produce meaningful work products	Core Proposition 1 Core Proposition 2 Core Proposition 3 Core Proposition 4
Using Assessment in Instruction	Criterion 6: Using multiple student data elements to modify instruction and improve student learning	Standard 2.2: Align assessments to curricula, use ongoing assessment and grading practices, and analyze information on student learning outcomes to adjust instructional decisions at the team and classroom levels	Core Proposition 1 Core Proposition 2 Core Proposition 3 Core Proposition 4

This mapping of the FFT dimensions to professional teaching standards in other contexts provides content-related evidence in support of the validity of FFT score inferences. Although the FFT dimensions are not identical to any of the teaching standards indicated here, there is clear overlap in the professional teaching standards in these three contexts and the FFT dimensions. As Danielson suggests in the FFT documentation, the practices on the FFT encompass those that are key to the teaching profession. The overlap between these dimensions and several professional teaching standards are evidence of the reasonableness of Danielson's claims.

2.3.2.2 Evidence based on response processes. As an instrument, the design of the FFT requires outside raters to score the practices of classroom teachers. This situation is obviously quite different from students answering items on a test. In this type of scenario, the Standards for Education and Psychological Testing (2014) suggest that

> relevant validity evidence includes the extent to which the processes of observers or judges are consistent with the intended interpretation of scores. For instance, if judges are expected to apply particular criteria in scoring test takers' performances, it is important to ascertain whether they are, in fact, applying the appropriate criteria and not being influenced by factors that are irrelevant to the intended interpretation. (pp. 15-16)

The researchers on the MET project paid special attention to the validity of rater scores when scoring videos with the particular protocols. First, raters participated in extensive training that included about 20-hours of hands-on work with each protocol. Training materials for the FFT included those provided previously in Figures 2.1 – 2.3. Recall that each of these training artifacts include detailed descriptions the dimensions of the FFT overall as well as explicit examples of teacher behaviors demonstrative of each scoring level. The Methods chapter of this dissertation describes the process of rater training in detail, but an important part of the training included a certification test. According to the Bill and Melinda Gates Foundation (2013), the

certification test included an activity in which raters scored videos previously assigned scores by project experts. That is, the project researchers considered these scores as "truth," or the scores teachers should receive on the teacher practices demonstrated in the training videos. Raters had to score these videos at a minimum level of agreement with expert scores. The criteria for passing the certification test varied by observation protocol. The FFT requirements included at least a 50% exact match of correct scores and no more than 25% of scores that were two or more off from the scores provided as "truth" (Kane & Staiger, 2012). Failing scores on the certification test resulted in exclusion from participation in the MET project.

Beyond the initial certification test, the researchers on the MET project established protocols to help ensure the validity of scoring after the initial rater certification. At the beginning of every rating "shift," raters began by scoring what the MET project referred to as calibration videos. Similar to the videos for the certification test, project experts assigned the calibration videos scores designated to represent "truth." Raters had to score the calibration videos at a pre-established (unspecified) level prior to moving on with the assigned videos for the session. In addition, "validity videos" made up 5% of the videos scored by each rater during their rating sessions. As with the previous certification videos, these validity videos had "true scores" attached to them, and members of the MET project team ensured that raters consistently applied the protocol rubrics on these validity videos. Finally, during each rating session, a scoring leader monitored the scores reported by each rater. As part of their responsibilities, scoring leaders double scored at least one video from each rater for the session. Double scoring safeguarded the validity of scores by session as well as identified those raters who might need additional training. Scoring leaders asked raters to redo the calibration scoring if necessary,

counseled raters if there were issues with the double scored videos, and even ended a rating shift if rater scores proved to be continuously problematic (Bill & Melinda Gates Foundation, 2013).

The proposed use of FFT scores is to identify differences in fundamental teacher behaviors. The practices put into place by the MET project researchers created a system of quality control on observation protocol scores to ensure that the application of protocols was consistent with designated "truth." Thus, rater scores serve as response process evidence of the validity of using FFT scores as a measure of key teacher practices.

2.3.2.3 Evidence based on internal structure. Another important consideration regarding the validity of FFT scores is that of the internal structure of the data. The proposed interpretation of FFT scores described at the onset of this validity argument is that FFT scores identify aspects of teacher responsibilities that improve student learning. There are multiple practices, but each is a piece of evidence about different behaviors in which teachers should engage. The Standards for Education and Psychological Testing (2014) suggest that validity evidence about the internal structure of scores should be consistent with the conceptual framework of the instrument.

As part of a study of the underlying factor structure of each of the protocols used in the MET project, McClellan, Donoghue, & Park (2013) investigated the factor structure of the FFT with an exploratory factor analysis (EFA). McClellan et al.'s study used data from the MET project, so, similar to the current project, it only included dimension scores from two domains on the FFT. The design of the FFT suggests that teacher behaviors related to classroom environment (Domain 2) and instruction (Domain 3) are distinct from one another. McClellan et al. began their study by comparing a two- and three-factor model. Their results supported a two-factor

structure because a third eigenvalue was smaller than one, and the third factor was not readily interpretable with respect to the FFT dimension loadings.

Next, McClellen et al. investigated the potential of a one-factor model as opposed to a two-factor model. This exploration occurred because only seven of the eight dimensions loaded on the appropriate factors. Specifically, Establishing a Culture for Learning (Domain 2) loaded on the same factor as all of the Instruction (Domain 3) dimensions. Additionally, McClellen et al.'s findings indicated that the proportion of explained variance only increased by about six percent when adding the second factor. Finally, the factors correlated at about 0.73 across both years of the project. Although McClellan et al. provide some evidence to suggest that the FFT has two factors that align with the two domains of the FFT used in the MET project, they ultimately suggest that the factors might not be structurally distinct given the high correlation between the two factors. Instead, the FFT may only have one underlying factor.

Preliminary exploration with the subset of MET project data used in this dissertation also included an EFA of the FFT data. Similar to McCllelan et al., the findings from that EFA indicated that all of the dimensions except for Establishing a Culture for Learning loaded as expected based on the FFT design. The correlation between the two factors was 0.71, which is also comparable to previous findings. However, unlike McClellen et al., the increase in the proportion of explained variance was just over 20%, as opposed to 6%. These somewhat differential findings suggested a need for confirmatory factor analyses (CFA).

The purpose of a CFA is to test hypotheses about the relationship between observed indicators (FFT dimension scores) and the latent factors they supposedly measure (FFT domains). The results of the EFA led to a CFA comparing two models: a model that assumed a two-factor structure based on the design of the FFT, and a more parsimonious, single-factor

model. The details regarding the CFA can be found in Appendix A. Table 2.4 shows the fit statistics for the two models. These results indicate that the two-factor model is a better fit for the data due to the higher comparative fit index (CFI) as well as lower root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR). Although the SRMR is not as low as the general rule of 0.05, the CFI and RMSEA meet the standard minimum thresholds. Further, the chi-square statistics and the statistical significance of the difference in chi-square statistics also indicate that the two-factor model is a better fit than the one-factor model (Hu & Bentler, 1999).

Table 2.4 CFA Model Fit Statistics

Model	DF	CFI	RMSEA	SRMR	Chi-Square	Chi-Square Diff	P-value
2-factor	43	0.987	0.065	0.062	1897.7	595.23	<0.01
1-factor	44	0.982	0.076	0.077	2615.1		

Although McClellen et al.'s analysis provides mixed results regarding the internal structure of the FFT, the subsequent EFA and CFA completed for the current project provide compelling evidence for a two-factor structure to the data. The CFA results specifically suggest that the FFT data supports the original design Danielson intended. This research serves as additional validity evidence regarding the use of FFT scores as Danielson defined and as used in this project.

Despite this two-factor structure, the longitudinal growth trajectories in this project use a composite score across all eight dimension-level scores. The more parsimonious score is appropriate for a first pass at understanding potential growth in teacher observation scores over time. If it is the case that there are notable changes in a general composite score, then those results provide evidence to suggest further work investigating change within domain as opposed

to across. Although identifying change in domain scores is a most logical second step, the second line of HLM analysis in this project focuses on dimension-specific growth trajectories because the conceptual framework suggests novice teachers vary regarding specific practices more than their experienced counterparts. This literature motivates an investigation into which (if any) dimension-level growth trajectories behave differently over time for novice versus experienced teachers.

2.3.2.4 Evidence based on relations to other variables. Finally, this section provides a brief review of three studies regarding the relationship between FFT scores and some student outcomes as a final source of validity evidence. The Standards for Education and Psychological Testing (2014) state that if intended score interpretation implies that the construct relates to some other variables, analyses of those relationships should be a part of the validity argument. Since the FFT explicitly states that scores from the protocol represent aspects of teacher's responsibilities that promote improved student learning it is important to provide evidence of this relationship.

In 2004, Milanowski conducted a study regarding the validity of FFT scores in Cincinnati Public Schools. Milanowski examined the relationship between teacher evaluation scores based on a modified version of the FFT to a value-added (VA) measure of student achievement. VA scores are a metric commonly used for teacher evaluation based on student achievement data. Student achievement scores, typically from state standardized tests, are aggregated up to the classroom level and used as a measure of teacher quality. Although there are numerous ways to estimate the scores, the general idea is that student achievement in a specific teacher's classroom is compared to the expected level of achievement if the students had been in an average teacher's

classroom instead. VA scores then provide evidence as to whether a specific teacher has "added value" to a student's achievement above and beyond what would be expected had the student been assigned to an "average" teacher (c.f. Aaronson, Barrow, & Sander, 2007; Briggs, 2012; Chetty, Friedman, & Rockoff, 2014; Gordon, Kane, & Staiger, 2006; Harris, 2009).

Milanowski's study included teacher evaluation scores for 212 teachers and state-level test scores for their respective students in the 2000-2001 and 2001-2002 school years. The sample included teachers in their first, third, fifth, or subsequent fifth (i.e. 10^{th}, 15^{th}, etc.) year of teaching, who taught in tested subjects, and who had more than three students. Teachers received scores on each of the four domains on the FFT. Milanowski summed across the domains to give teachers a total score for each observation and used a two-step regression to identify teacher VA scores. The model for estimating VA scores included controls for prior achievement, sex, free-and-reduced-price lunch status, race/ethnicity, special education status, and the number of days of student enrollment. Milanowski found the rank-order correlations between VA and teacher observation scores. The weighted averages across grades indicated a correlation of 0.21 in reading and science and 0.30 in math. These findings suggest that teacher observation scores have a moderate relationship with value-added estimates of teacher effectiveness. In other words, Milanowski's work provides evidence of a relationship between teacher observation scores and higher achievement for students aggregated to the teacher level.

In a similar study Kimball, White, Milanowski, & Borman (2004) conducted an investigation of the validity of FFT scores in Washoe County using HLM techniques. The study included 328 third- through fifth-grade teachers in the 2000 – 2001 and 2001 – 2002 school years. The teachers were those in tested subjects and in their first or second year of teaching or a "major evaluation" phase of their careers. Every three years after their first two years, teachers

cycle through major and minor evaluation phases. During the major evaluation phase, observations occur three times a year on two FFT domains in one year and then the other two domains in the following year. After this cycle, teachers advance to a minor evaluation in which ratings occur in only one domain. Teachers in this minor evaluation are the only ones not included in the sample. The model for estimating VA scores in this study included controls for prior achievement, sex, free-and-reduced-price lunch status, race/ethnicity, special education status, teacher education, teacher experience, and academic calendar (i.e. year-round vs. traditional calendar).

This analysis resulted in a correlation between teacher value-added and teacher observation scores ranging from about 0.10 to 0.40 depending on the grade and subject-area. Third-grade reading and math both correlated to observation scores at around 0.10. Fourth-grade correlations varied by subject. Reading correlated with observation scores at about 0.28 while math correlated at about 0.07. Finally, fifth-grade reading also correlated to observation scores at about 0.28 and math at approximately 0.37. Kimball et al.'s findings indicate a weak to moderate relationship between teacher observation scores and student achievement. The strongest relationships between teacher observation scores and student achievement occurred in fourth grade reading and both subjects in fifth-grade. The current study includes both subjects in fourth and fifth grade, so even though Kimball et al.'s results provide moderate evidence of the relationship between FFT scores and student achievement, the lower correlations for fourth-grade math indicate mixed results.

In addition to this prior work regarding the validity of the FFT, researchers on the MET project also considered the validity of scores from the observation protocols within the project itself. Kane & Staiger (2012) investigated the association between observation scores and value-

added estimates based on longitudinal student achievement data. Specifically, they estimated VA scores that controlled for prior performance, similar demographic characteristics[9] (i.e. age, race/ethnicity, gender, free or reduced price lunch status, English Language Learner status, Special Education status, and Gifted and Talented status), and similar characteristics aggregated to the classroom level. The correlations between teachers' value-added scores and observation scores from the FFT ranged from 0.11 in English language arts to 0.18 in math. Although these correlations are modest, Kane & Staiger argue that the ultimate goal of classroom observations is to help teachers improve their student outcomes. Since these relationships were positive, despite marginal magnitude, they provide these associations as evidence of the validity of the observation scores in the MET project. Additionally, as will be discussed further later in this dissertation, there is limited variability in observation scores. This limitation precludes much stronger correlation between VA and observation scores.

2.3.2.5 Validity evidence summarized. Taken alone, the studies presented regarding relationship between FFT scores and student achievement provide mixed evidence for the validity of FFT scores. However, The Standards for Education and Psychological Testing (2014) state that a strong validity argument "integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (p. 21). The previous sections (2.3.2.1 – 2.3.2.4) describe

[9] Demographics varied by district. Charlotte-Mecklenburg includes race, ELL status, age, gender, special education, & gifted status; Dallas includes race, ELL, age, gender, special education, free or reduced-price lunch status; Denver includes race, age, ELL, free or reduced-price lunch status, gender, and gifted status; Hillsborough includes race, ELL, age, special education, gifted status, and free or reduced-price lunch status; Memphis includes race, ELL, free or reduced-price lunch, gender, gifted status, and special education; and New York City includes race, ELL, gender, special education, and free or reduced-price lunch status (Kane & Staiger 2012).

multiple sources of evidence of the validity of FFT scores for providing information about teacher practices that are key to the profession. Taken together with the mixed results regarding the relationship between FFT scores and student achievement, there is relatively consistent evidence from a variety of sources of the validity of FFT scores as indicators of key behaviors of the teaching profession.

2.3.3 Reliability of the FFT. The next section considers the reliability of FFT scores in the MET project. The Standards for Education and Psychological Testing (2014) state that reliability in a general sense is the consistency of scores across replications of a testing procedure or, in this case, the application of an observation protocol. More specifically, the reliability of scores "depends on how much the scores vary across replications of the [observation] procedure" (p. 33). In other words, the protocol scores assigned on any given occasion should be consistent regardless of factors such as rater or occasion. It is important that variation in observation scores consistently reflect variation in teacher practices and not idiosyncrasies due to rater or lesson involved in the observation. Thus, investigation of the reliability of observation scores must consider the kinds of variability that can affect those scores.

In their study of the reliability of observation scores, Kane & Staiger (2012) studied a small subset of teachers in the MET project who received scores from more than one rater. Within this subset, they identified the degree to which scores varied from teacher to teacher, section to section, lesson to lesson, and rater to rater. This analysis included a decomposition of the total variance in observation scores into variance components for both total FFT score as well as dimension-level scores.

The variance components in Table 2.5 indicate variability in FFT scores for a single lesson with a single rater. For example, the variance in FFT scores due to teacher for a single lesson and single rater range from 18 – 33% across specific dimensions and 37% for overall scores. In other words, the reliability of FFT scores based on one lesson and one rater is 0.37. Broadly, this is a relatively low reliability estimate. However, Kane & Staiger's results are consistent with what is seen in other literature (reliability ranging from about 0.3 to 05) regarding one observation with one rater (c.f. Hill et al., 2012; Ho & Kane, 2013). Kane & Staiger used these variance components to estimate how reliability changes with two lessons and one rater as well as four lessons and one rater. They found that increasing to two and four lessons and averaging scores over those lessons increased reliability to 0.53 and 0.67 respectively for the overall lesson. Additionally, dimension-specific score reliabilities ranged from 0.38 to 0.60 for four lessons, each with a different rater.

Table 2.5 Variance Decomposition and Implied Reliability for the FFT

	Percentage of Variance					Implied Reliability		
Dimension	Teacher	Section	Lesson	Rater	Residual	1 lesson 1 rater	2 lessons 1 rater each	4 lessons 1 rater each
Total	37	4	10	6	43	0.37	0.53	0.67
CERR	30	3	8	7	53	0.30	0.45	0.60
ECL	25	0	10	7	57	0.25	0.40	0.58
MCP	24	6	0	7	62	0.24	0.37	0.51
MSB	33	8	1	3	54	0.33	0.47	0.59
CS	21	1	2	8	68	0.21	0.34	0.50
USDT	15	4	12	6	62	0.15	0.25	0.38
ESL	20	3	12	6	59	0.20	0.33	0.47
UAI	18	0	3	9	70	0.18	0.31	0.47

Creating an environment of respect and rapport (CERR)
Establishing a culture for learning (ECL)
Managing classroom procedures (MCP)
Managing student behavior (MSB)
Communicating with students (CS)
Using questioning and discussion techniques (USDT)
Engaging students in learning (ESL)
Using assessment in instruction (UAI)

Adapted from Gathering Feedback for Teaching (p. 35), by Kane & Staiger (2012)

The reliabilities in the Kane & Staiger study derive from three different measurement designs: one lesson with one rater; two lessons, each with a different rater; and four lessons, each with a different rater. Their findings suggest that the reliability of FFT scores for only one occasion with only one rater is 0.37 for total score, but reaches 0.67 with four lessons, each with a different rater. As such, the reliability of only one lesson with one rater is rather limited.

It is important to pause at this point to consider two different ways of considering growth in teacher practices as they have different implications for the way we think about reliability in this project. If we assume, for instance, that growth in teacher behaviors only occurs between years, then observation occasions within year are interchangeable. Further, any difference in observation scores within a year are due to measurement error rather than salient changes in teacher practices. Under this assumption, one purpose of additional occasions within year is to make FFT scores more reliable. As is done in the aforementioned Kane & Staiger (2012) study, researchers often increase the reliability of these "status scores" by averaging observation scores across multiple occasions. When this is done, researchers ignore (or conflate) the occasion of measurement with lesson. Any differences that exist in observation scores across occasions that might be due to salient changes in teacher practice are averaged away as measurement error. When it is assumed that there is no meaningful time trend in observation scores, it makes sense to use additional observations in this way so that we attain the most reliable status scores.

In contrast, if we assume that growth in teacher practices occurs both between *and* within years, we assume that there is important information about a time trend in teacher practices present at every observation occasion. That is, differences in observation scores within year are not completely due to measurement error, but rather at least a portion of those differences is

signal of teachers changing their practices in meaningful ways. If we consider growth in this way, it does not make sense to average observation scores over time because information about if and the extent to which teachers change their behaviors is lost.

The distinction between these two different ways of thinking about growth in teacher practices is important when considering Kane & Staiger's findings in light of the current study. Kane & Staiger find that the reliability of FFT status scores for a single occasion with a single rater is 0.37 for total FFT scores and about 0.25 for the dimension-specific scores. Although this reliability can reach much higher levels if time is treated as a source of error variance, growth trajectories that use each occasion as a separate indicator of signal in teacher practices rely on the reliability of a single lesson with a single rater in the MET dataset. Thus, Kane & Staiger's findings suggest that the longitudinal estimates of growth in teacher practices in this study are based on FFT scores with reliability of 0.37. Although this is conventionally rather low, it is important to note that this falls within the range of reliabilities typically reported for VA scores, another measure of teacher quality (Kane & Staiger, 2012; McCaffrey et al., 2009).

Since the current study focuses on detecting the trend over time both within and across years, I begin with a collection of FFT scores and fit a linear growth trajectory for each teacher across all occasions. The slope for each teacher is an estimate for the teacher-specific growth over time in observation scores. The reliability of this slope estimate is the focus of the second major line of inquiry in this dissertation. However, I also include a brief analysis that assumes growth in teacher practices only occurs across years rather than within *and* across. In this approach, all of the scores within one year are averaged together. This averaging does two things. First, it makes each of the two status scores more reliable indicators of teacher practices within year, so the basis of the growth estimates are more reliable. According to Kane &

Staiger's results, growth in this approach is based off of individual scores with reliabilities of about 0.67 as opposed to 0.37. Second, it assumes there is no meaningful growth in teacher behaviors within year. Thus, growth is calculated with a gain score between the two years. The reliability of these gain scores is compared to the reliability of the growth parameter estimates[10] to provide information regarding how choices about the way we think about how change in teacher practices occurs affects the reliability of the resulting growth estimates.

2.4 Current Project

Although educational researchers have not historically had the opportunity to investigate change in teacher practices over time as measured by observation scores, it is entirely possible and even preferable for teachers to improve in their practices over the course of their careers. Instructional coaches on campuses, teams of teachers planning and reflecting on lessons together, and district practices of providing professional development experiences for their employees are evidence of this preference. The current research on teacher change includes a discussion of the ways in which professional development opportunities influence teacher behavior as well as how teacher change manifests itself in teacher beliefs or in changes in student outcomes. However, there is yet to be research regarding change over time in teacher practices as measured by observation scores.

---

[10] Although both gain scores and growth parameter slope estimates are technically growth parameter estimates, I use the term "growth parameter estimates" to refer to the growth estimates derived from using each occasion as an individual source of data in the estimation and gain score to refer to estimates derived from scores averaged within year.

This dissertation thus serves as a methodological model for estimating growth in teacher observation scores. The evidence for the validity and reliability of FFT scores presented in this chapter suggest that the FFT is a reasonably appropriate observation protocol with which to try estimating growth over time in teacher practices. Further, the investigation of how the details of observation systems might influence the reliability of these growth estimates is of practical interest. Developing a method to quantify growth would be helpful to schools and districts as this sort of information can be used to identify teacher leaders as well as more careful planning of specific professional development activities for teachers or even as one of multiple measures of teacher quality. The next chapter provides a detailed description of the subset of data from the MET project used to develop longitudinal growth trajectories based on scores from the FFT.

3.0 Data

This chapter describes the subset of data from the MET project used in this dissertation. Section 3.1 includes a brief account of the MET project generally, and describes the video collection process employed in the MET project. Next, Section 3.2 explains the process by which the MET project recruited and trained raters, and includes discussion regarding the scoring process with the FFT. There is also a brief discussion of the distribution of scores across dimensions of the FFT. The chapter closes in Section 3.3 with a presentation of the data structure and a discussion of how the specific data structure in the MET project lends itself to developing individual growth trajectories.

3.1 Measure of Effective Teaching Project

The data used in this work comes from the MET project (Bill and Melinda Gates Foundation, 2013). The Bill and Melinda Gates Foundation funded the project to investigate multiple measures of teacher quality, including observation and test scores as well as student perception surveys. The full study includes information on approximately 3,000 fourth- through ninth-grade teachers at over 300 schools in six US school districts: Charlotte-Mecklenberg, North Carolina; Dallas, Texas; Denver, Colorado; Hillsborough County, Florida; New York City, New York; and Memphis, Tennessee (Cantrell & Kane, 2013). Data collection occurred during the 2009-2010 and 2010-2011 school years. The design of the MET project required teachers to submit four video lessons for each of the two years of the study for a total of eight

videos per teacher per subject. Each of these videos received scores from multiple observation protocols, including the Charlotte Danielson Framework for Teaching (FFT).

3.1.1 Data cleaning. The current study uses information from the FFT for all 4th- through 9th-grade math and ELA teachers in the MET data set. Of the original 2,741 teachers involved in the project at large, 1,569 received FFT scores, but only about 953 teachers had dates associated with their videos (for reasons that are not explained in the MET data set). In addition, approximately half of these 953 teachers do not have values for the variable indicating their years of experience. This is because two of the districts did not provide information for this variable to the MET researchers (Kane & Staiger, 2012). Since years of experience is necessary to examine differential growth rates based on a teacher's status as novice or experienced, the dataset for this project does not include cases in which there is no information regarding a teacher's years of experience. As a result, the final dataset used to conduct the HLM analysis includes 458 teachers who provide 3372 unique videos. Figure 3.1 illustrates the funneling of this data from the complete MET data to the final sample.

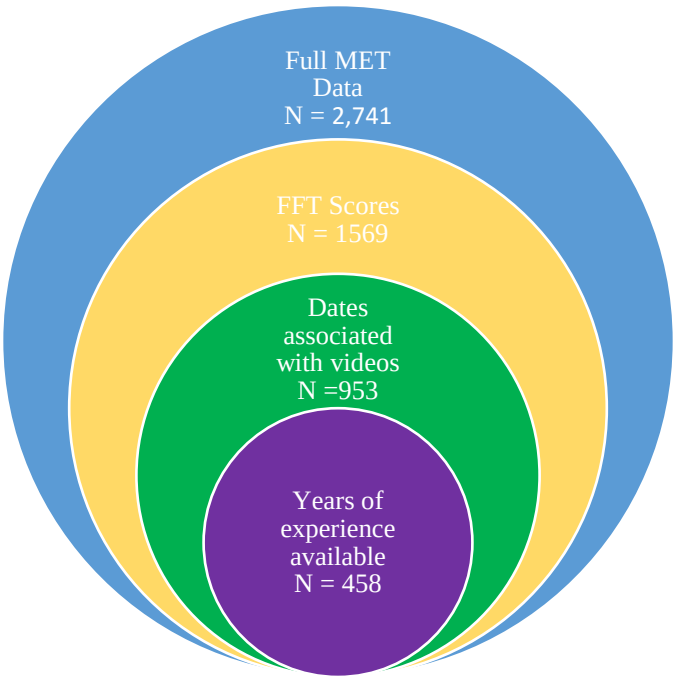


Figure 3.1 Number of teachers available in each subsequent subset of the MET data

3.1.2 Demographics. The information in Table 3.1 provides a summary of the demographic information available for the teachers in the purple, green, and yellow sections of Figure 3.1. The table allows for comparisons between these groups of teachers and provides cursory information on whether the teachers used in the current project are noticeably different from teachers for whom experience or date information was not available.

Table 3.1 Teacher Demographics Percentages (Counts)

	Experience Available (Purple)	Experience Unavailable (Green)	Dates Unavailable (Yellow)
Male	22 (100)	16 (77)	17 (103)
White	65 (296)	53 (260)	53 (328)
Black	25 (113)	36 (177)	36 (224)
Masters +	23 (107)	33 (161)	27 (164)
Elementary (Grades 4-5)	26 (117)	47 (235)	40 (245)
Secondary (Grades 6-9)	74 (341)	53 (260)	60 (371)
Novice	16 (72)	NA	6 (36)*
N	100 (458)	100 (494)	100 (616)

*Note: 65% of teachers without dates were also without experience information

The demographics presented in Table 3.1 suggest that the group of teachers in the final analysis are more likely to be male and white and less likely to have a master's degree or higher than the other teachers eliminated from the sample due to missing information. Further, about half of the teachers without information about experience and 40% of those without dates attached to their videos taught elementary school, while only about one quarter of the teachers for whom information about experience and video date was available taught elementary school. This means that most teachers in the study would have been required to submit only four videos per year. Thus, although it was possible to submit 16 videos per teacher, the average number of videos per teacher will be much closer to 8 than 16 due to the grade-level make-up of the teachers in the sample. All of these differences in teacher demographics indicate limitations in the generalizability of the findings in this study that are discussed at the close of this dissertation.

3.1.3 Video collection. The MET project design expected teachers to submit four classroom videos per subject between February and June of 2010 for Year 1. In Year 2, researchers expected four more videos per subject between October 2010 and June 2011. Project researchers encouraged teachers to spread the video recordings over each span of time within each year to ensure that the recordings were more representative of instruction than a series of closely timed lessons. However, there were no constraints set on the amount of time required between each video occasion (Bill and Melinda Gates Foundation, 2013).

In addition to variation in time between the video occasions, the researchers asked teachers to vary the topics covered in the videos. Half of the videos within subject within each year were to be focused on a set of "focal" topics which were pre-determined by the MET researchers (e.g. "Making inferences/Questioning", "Personal Narrative", "Operations on

Rational Numbers", and "Multiplication and division of fractions or decimals"). Teachers chose the topics for the other half of the videos.

Teachers were trained and responsible for video recording as well as uploading the video to a secure website. Two cameras collected the data for each occasion. One specially designed camera provided a 360-degree view of the classroom, and the other camera focused exclusively on what the teacher wrote on the board. Two wireless microphones captured the teacher's voice as well as student voices. After each video collection, the teachers uploaded the separate video and audio files, and a researcher combined them all into one video. Teachers then reviewed the videos to check for accuracy, uploaded student work, lesson plans, and written comments on the lesson. After video submission, they were ready for scoring by MET raters with the appropriate observation protocols (Bill and Melinda Gates Foundation, 2013).

3.2 Rater Recruitment, Training, and Video Scoring

The Literature Review chapter provides details regarding the FFT protocol development as well as the validity and reliability of scores from the FFT. This section provides the relevant information regarding the rater recruitment, training, and scoring with the FFT in the MET project.

Educational Testing Services (ETS) recruited and managed the 902 MET project raters (Kane & Staiger, 2012). Recruitment occurred through various avenues including postings on the ETS website and the websites of professional organizations such as the National Council of Teachers of Mathematics. Additionally, ETS sent emails to previous and current raters from their own organization as well as put postings on Facebook. Rater recruitment through postings on the

ETS and other professional websites were most successful. All raters had at least a bachelor's degree, and more than 75% of the raters had six or more years of teaching experience. Some raters were in teacher preparation programs and had yet to enter the classroom as the teacher of record. Specifically, 9% of the FFT raters had no previous teaching experience (Kane & Staiger, 2012).

The MET project required online training and certification for all raters. The training modules for the FFT were initially developed by the Danielson Group, Charlotte Danielson's organization that manages work related to the FFT, and the Teachscape Corporation, an organization for teacher professional development. Each training was self-directed and lasted about 20 hours. The trainings included four main sections: (1) using the web interface for scoring, (2) eliminating bias from scoring, (3) understanding the purpose and method for the FFT, and (4) scoring for each dimension on the FFT. All trainings included video examples for each level of scoring for each element of a protocol. Additionally, the trainings included practice scoring and feedback from trainers. Following completion of the training, potential raters were required to pass an initial certification test. The test required potential raters to score the certification test videos at a minimum level of agreement with expert scores. The FFT requirements included at least a 50% exact match of correct scores and no more than 25% of scores that were two or more off from the scores provided as "truth" (Kane & Staiger, 2012). If potential raters failed this test on the first attempt, they reviewed the training materials one more time and then re-took the certification test. If potential raters failed the certification test a second time, they became ineligible to score for the MET project. Kane & Staiger (2012) indicate that 83% of potential raters passed the FFT certification test.

During the second year of the MET project, teachers received rosters of students that were randomly assigned within-grade within each school. The researchers on the MET project referred to these grade-level groups of classrooms as "randomized blocks." Every classroom within a randomized block received scores from the same group of raters, and each rater scored one video for each teacher in a block. Additionally, raters only scored a single video for a given teacher in a given year. That is, the most a single rater scored a single teacher was twice over both years of the study. However, no procedures ensured that a rater scored a specific teacher twice in the study. Further, only about 5% of videos received scores from two raters. All of the rest of the videos only received scores from one rater per protocol. Double-scored videos were selected at random, and raters did not know when they scored these videos. In addition, raters could not score videos from teachers they knew or teachers in districts where they had worked or had any affiliation. The method of assignment of rater to video precludes any analysis investigating the variation in scores due to rater since nearly every score assigned to a given teacher came from a different rater and only a very small number of the videos received scores from multiple raters[11]. Of the 3372 unique videos included in the data for this project (those from the purple circle in Figure 3.1), 181 (or 5%) received double scores. For each video that was double-scored, one randomly selected set of scores appears in the final data set for estimating growth trajectories in this project.

Recall from the Literature Review chapter of this dissertation that the FFT includes four domains. Of those four, two were the focus of the MET project context: Classroom Environment

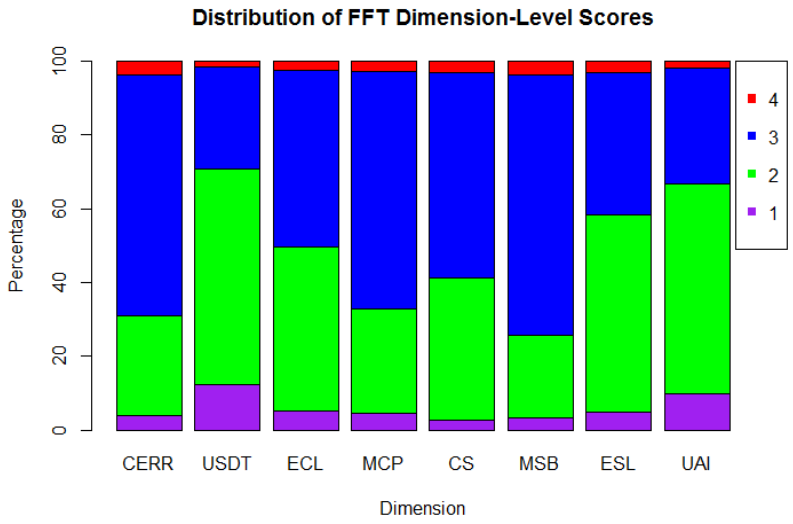
[11] Due to this data limitation, Ho & Kane (2013) conducted a follow-up study in Hillsborough County, FL where multiple raters scored each lesson in order to disentangle potential rater effects. This study is discussed in the Results chapter to help interpret the reliability of the growth parameter estimates in the current project.

and Instruction. Each of these domains includes four dimensions, for a total of eight individual scores per video. Scoring for each dimension ranges from one to four, so total scores range from eight to thirty-two. Appendix B presents the rubrics as represented in the MET project data user's guide for all eight dimensions, but Table 3.2 provides one dimension as an example.

Table 3.2 Scoring Rubric for Managing Classroom Procedures

Score	Classification
1	Unsatisfactory—instructional time is lost, inefficient classroom routines
2	Basic—some instructional time is lost, partially effective classroom routines
3	Proficient—little loss of instructional time, consistently follow established classroom routines
4	Distinguished—instructional time is maximized, students contribute to management and classroom routines

Table 3.2 illustrates the scoring rubric used for the third dimension of the Classroom Environment domain, managing classroom procedures. Raters used rubrics similar to the one provided in Table 3.2 to score each of the eight dimensions on the FFT for each video in the MET project. During scoring, raters viewed the first fifteen minutes of each video and then skipped to viewing minutes twenty-five through thirty-five. Figure 3.2 illustrates the distribution of scores for each of the eight dimensions on the FFT across all occasions.



Creating an environment of respect and rapport (CERR)
Establishing a culture for learning (ECL)
Managing classroom procedures (MCP)
Managing student behavior (MSB)
Communicating with students (CS)
Using questioning and discussion techniques (USDT)
Engaging students in learning (ESL)
Using assessment in instruction (UAI)

Figure 3.2 Distribution of FFT dimension-level scores

It is obvious from Figure 3.2 that the vast majority of teachers received scores of two or three for all of the dimensions of the FFT. Another way of considering these scores is as the average dimension-level score per occasion. In other words, for each occasion, calculate the average of all eight dimension-specific scores. Considering overall score within an occasion in this way allows the reader to think consistently within the range of one to four, regardless of the outcome of interest (dimension-specific or mean overall score) as opposed to considering total score (ranging from eight to thirty-two). Figure 3.3 presents the distribution of these average scores across occasions.

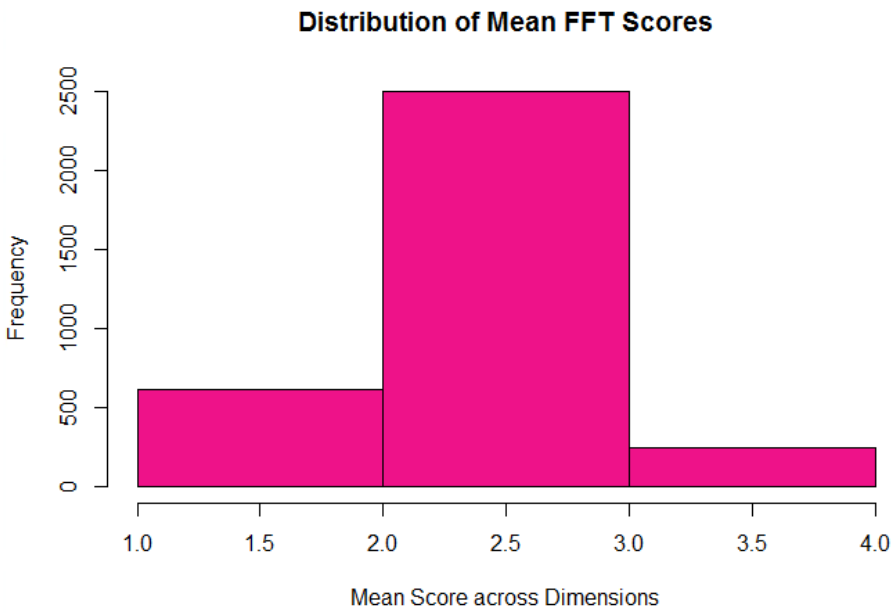


Figure 3.3 Histogram of mean FFT scores across occasions

This figure gives the frequency (y-axis) of each video's average FFT dimension-level score (x-axis). Each dimension received a score between one and four, so the average of the eight scores also ranges from one to four. The bars in this histogram are inclusive of the right-hand value. This means that the first bar includes all videos with an average score ranging from one to two. The middle bar illustrates all videos with a mean score greater than two and less than or equal to three, and the final bar indicates the number of videos with a mean score greater than three. The sample used in this project includes about 3,300 videos, and about 2,500 of them received mean scores between two and three. The mean average dimension-level scores across all teachers and occasions is 2.5 and the standard deviation is 0.47. Consistent with Figure 3.2, the distribution in Figure 3.3 suggests that very few teachers receive scores of one or four in the MET project. However, presenting these scores across all occasions masks potential time trends. One additionally helpful way of considering these scores is the average for each occasion, separately.

Table 3.3 provides the mean FFT scores for the first eight occasions for all teachers with at least eight occasions in the MET project dataset. It is important to note that the timing of each of these occasions is different for every teacher. For example, one teacher may have her first occasion during the first week of the MET project, and another might have his first occasion two months later. The scores in Table 3.3 average across every teacher's individual occasions regardless of when they occurred during the calendar year. Reporting mean FFT scores in this way foreshadows the results to come in Chapter 5. That is, although there is some variance in mean FFT scores over occasions, there is little evidence to detect a trend in scores over time.

Table 3.3 Mean (SD) FFT Score by Occasion

	1	2	3	4	5	6	7	8
FFT Score	2.48 (0.45)	2.53 (0.46)	2.50 (0.44)	2.45 (0.46)	2.51 (0.47)	2.49 (0.50)	2.51 (0.48)	2.48 (0.49)
N = 292								

3.3 Data Structure

A simplified example of the structure of the general data used for this analysis is in Table 3.4[12]. The organization is by person-period, or a data in long format with a separate line in the data for each teacher on each occasion. Each line includes not only identification variables such as district, school, teacher, and video identification, of which only teacher is illustrated here, but also the separate FFT dimension-level scores and the average of the dimension-level scores for each occasion. There are eight FFT dimension-specific scores, but only two are presented for

[12] The interested reader can find a complete list of all the available variables and how they were compiled in Appendix C.

illustrative purposes. There are also teacher and classroom demographic variables such as a teacher's sex and the percent of students in a classroom identified as English Language Learners (ELL). For the sake of space, Table 3.4 does not include all of these demographic variables. Rather, only an indicator for teacher experience labeled as Novice (coded as "1" for less than three years of experience in the first year of the MET project; coded as "0" for three or more years of experience) appears as it is most relevant to the HLM models explained later in the Methods chapter. Next is a column that simply numbers the subsequent occasions for each teacher. Finally, there are also three indicators of time. First is the raw date variable indicating the date of the video recording. Following the date variable is a variable indicating day. This variable simply numbers the days from the beginning to the end of the MET project and translates the date variable into the number of days between the beginning of the project and the date of the video. Something to note is that this numbering does not include the days in the summer between the two years of the MET project. In other words, the date counting numbered subsequently from the last day of the first year to the first date of the second year. This choice eliminates the space of time over the summer[13]. The final variable is the Week variable. This is the indicator of time used in the longitudinal growth trajectories. Week was chosen over day because growth in a teacher's practices by day seemed too narrow for logical interpretation while month potentially too wide. The Week variable calculation consisted of simply dividing the Day variable by seven and rounding to the closest whole number.

---

[13] It is necessary to make a choice between counting time subsequently across years, including summer, or counting time subsequently across years, eliminating summer. I investigated both approaches, and mean reliability results were similar regardless of inclusion or exclusion of summer. Differences in magnitudes occurred at the hundred-thousandths place in the variance components. I chose to eliminate summer because this was a space of time during which no teachers would be observed. Thus, no differences in behavior could be observed.

Consider Teacher 1 in Table 3.4. She submitted her first video on 2010-04-20. This date is 78 days from 2010-02-01 and occurred in the 11th week. Similarly, Teacher 1's final video recording took place on 2011-03-29, which is toward the end of the second year of the MET project. Thus, the Week value for this date is 43.

Table 3.4 Simplified General Data Structure

ID	Year	Mean Dim Score	CERR	…	UAI	Novice	Occasion	Date	Day	Week
1	2010	2.6	3	…	3	0	1	2010-04-20	78	11
1	2010	2.4	2	…	2	0	2	2010-05-04	92	13
1	2010	2.9	3	…	3	0	3	2010-05-11	99	14
….	….	….	….	…	…	….	…	….	….	….
1	2011	2.6	3	…	2	0	6	2011-03-28	420	43
1	2011	3.0	3	…	3	0	7	2011-03-29	421	43
….	….	….	….	…		….	…	….	….	….
85	2010	2.3	2	…	2	1	1	2010-04-26	84	12
85	2010	3.0	3	…	2	1	2	2010-04-27	85	12
85	2010	3.0	3	…	3	1	3	2010-04-28	86	12
….	….	….	….	…	…	….	…	….	….	….
85	2010	2.0	2	…	2	1	7	2010-04-29	87	12
85	2010	2.5	3	…	2	1	8	2010-05-03	91	13

Note: CERR = Creating an Environment of Respect and Rapport; UAI = Using Assessment in Instruction

Two notable features of this dataset is that it is both "time-unstructured" and "unbalanced" (Singer & Willett, 2003). A longitudinal dataset is time-unstructured if the timing of the subsequent occasions is not uniform across individuals. Although teachers in the MET project were asked to submit four videos for each subject over each of the two years of the study, for a potential total of sixteen videos per teacher, the spacing of these videos was not standardized. That is, each individual has his/her own schedule of observation occasions. For example, note that Teacher 1 has seven total videos. Of the seven, only five appear in Table 3.4. Three occurred in the first year of the MET project and two occurred in the second year. The

standard deviation of the Week variable for this teacher is 11.6. Conversely, Teacher 85 has a total of eight videos, and all of those are from the first year of the MET project. The standard deviation of the Week variable for Teacher 85 is 0.35. Thus, this data collection is time-unstructured, and these two example teachers indicate how variable the spread of occasions across teachers is in the MET project.

The second notable feature of the dataset is that it is unbalanced, or all of the teachers do not have the same number of video occasions in the dataset. The data in Table 3.4 indicate that Teacher 1 received FFT scores on seven unique occasions while Teacher 85 received eight scores. Although it was possible for a number of teachers to submit 16 videos to the MET project, the number of teachers who actually did this is rather small. Figure 3.4 illustrates the percentage of teachers with the corresponding number of occasions. For example, 100% of the teachers have one video in the dataset. However, after one occasion, the cumulative percentage of teachers with a given number of occasions drops steadily until eight occasions. Then there is a drop to about 10% of teachers with nine occasions. This percentage continues to drop through sixteen occasions. The design of the MET project called for those elementary teachers who instructed the same group of students in both math and ELA to submit four videos per subject per year, for a total of sixteen videos. Elementary teachers made up about 25% of the full sample of teachers used in this study. Thus, it is not expected that the percentage of teachers to have nine or more occasions to exceed 25%. However, less than 10% of teachers submitted more than eight videos, so fewer than half of the elementary teachers submitted the full number of videos requested of them in the MET project design. Despite this high level of attrition, Singer & Willett (2003) note that data need not be balanced across individuals for developing growth

trajectories. Instead, teachers who vary with respect to the number of occasions provide examples of unique measurement designs not originally intended by the MET researchers.

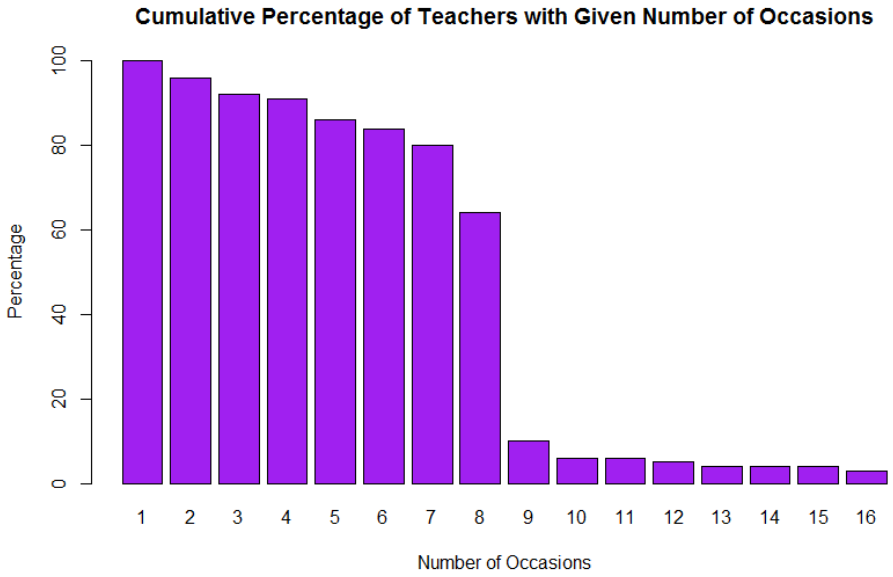


Figure 3.4 Distribution of teachers for each occasion

The differences in number and spacing of occasions illustrates what is meant by different measurement designs represented in the MET project. A local education agency designing an observation system must make choices regarding the number and spacing of observation occasions that should occur[14]. As mentioned previously and discussed in greater detail in later chapters of this dissertation, the choice of number and spacing of occasions has implications for the reliability of estimated growth parameters. Each teacher in the MET project thus represents a different measurement design that a local education agency could take up as part of its

[14] Number of raters for a given occasion is also a part of each measurement design, but all of the designs in the MET project employed only one rater per lesson.

observation system. However, these teachers do not represent an exhaustive list of potential measurement designs.

Before moving on to a description of the methods used in this project, this chapter closes with an illustration of the two datasets used to complete the HLM analysis. The data in Table 3.4 is the general dataset used to provide descriptive statistics and to get a general feel for the data. From this dataset, came two subsets of data, each related to a different level of the multi-level model explained in the next chapter. The two levels (discussed further in the Methods chapter) are at the occasion level and the teacher level. The data necessary to complete the level-one (occasion) analysis appears in Table 3.5.

Table 3.5 Level-One Data Structure

Teacher ID	Mean Dim Score	CERR	…	UAI	Week
1	2.6	3	…	3	11
1	2.4	2	…	2	13
1	2.9	3	…	3	14
….	….	….	…	…	….
1	2.6	3	…	2	60
1	3.0	3	…	3	60
….	….	….	…		….
85	2.3	2	…	2	12
85	3.0	3	…	2	12
85	3.0	3	…	3	12
….	….	….	…	…	….
85	2.0	2	…	2	12
85	2.5	3	…	2	13

Note: CERR = Creating an Environment of Respect and Rapport;
UAI = Using Assessment in Instruction

Table 3.5 depicts the subsample of data for the level-1 HLM analysis. The data include mean FFT score across dimensions for each occasion as well as each of the dimension-specific scores for every occasion (although only two appear here). Next is the Week variable described

above. These data are first used to model trends in growth generally and then identify if that growth is different for novice teachers for the mean FFT scores within occasion. After this first analysis, the project moves on to consider each of the dimension-specific scores as the outcome of interest to model change over time for each dimension and identify if those trends differ for novice as opposed to experienced teachers. This analysis is driven by the conceptual framework, which suggests that novice teachers' may experience steeper growth for some practices than others. If it is the case that dimension-specific variability is masked in the mean-score analysis, it is important information for individuals wishing to apply the methodology in this project to a new context.

The data for the level-two (teacher) model appear in Table 3.6. These data simply include the Teacher ID linking variable and the teacher-level experience predictor, Novice. Although these are also a subset of columns from that originally presented in Table 3.4, there is only one line per teacher as the number of occasions are not pertinent when identifying teacher-level variables. As noted earlier, Novice is coded one if a teacher is in her first two years of teaching at the beginning of the MET project and zero otherwise.

Table 3.6 Level-Two Data Structure

Teacher ID	Novice
1	0
…	…
85	1

The next chapter describes the specific methods used in this dissertation. It begins with the details of the Hierarchical Linear Models used in the current context, and describes their application for both overall scores and dimension-specific scores. The chapter then provides a

discussion of the reliability of the growth estimates in this project before considering how reliability might increase in other contexts.

4.0 Methods

The theoretical framework provided in Chapter 1 suggests that all teachers may change their practices, regardless of where they are in their careers, and teachers in their first two years of teaching tend to have more dynamic practices than more experienced teachers. Further, prior qualitative research suggests that novice teachers are particularly more variable on practices such as managing student behavior or managing classroom procedures. Thus, the methods used in this dissertation aim to answer three research questions:

1) to what extent do teacher observation scores change over the course of the approximately two years of the MET project?
2) are there significant differences in the rates of growth on observation scores for novice versus experienced teachers?
3) what conditions yield the most reliable estimates of growth over time in teacher observation scores?

This chapter begins in Section 4.1 with a presentation of the specific models used in the current study. Next is an explanation of the investigation of the reliability of the growth parameters in Section 4.2. The chapter closes in Section 4.3 with a brief summary and concluding remarks regarding the methods used in this project.

4.1 Hierarchical Linear Models of Change over Time

A variety of disciplines including sociology, education, biometrics, econometrics, health, social work, and business use hierarchical linear modeling (HLM). This widespread usage is largely due to the prevalence of nested data. However, the adoption by such a large variety of fields lead to an equally varied collection of terms for the methodology. HLM is known by several other names including multilevel-, mixed level-, mixed linear-, mixed effects-, random effects-, random coefficient (regression)-, and (complex) covariance-components-modeling. Each of these labels denote the same regression technique that is HLM (Raudenbush & Bryk, 2002; Woltman, Feldstain, MacKay, & Rocchi, 2012). Similar to Raundenbush and Bryk (2002), the current study uses the term HLM as this name demonstrates the greatest strength of the method—it provides a way to linearly model hierarchical, or nested, data. HLM provides information about the relationships both within and between the hierarchical levels in the data and makes it an ideal approach for attributing variance from the dependent variable of interest among independent variables at different levels of the data.

This dissertation project relies on three successive HLMs to provide information regarding the growth trajectories for teachers in the sample and then to see if those trajectories differ for novice teachers. These models are named 1) the unconditional means model, 2) the unconditional growth model, and 3) the novice model. The study includes nine runs of each model. First for mean dimension-level FFT score, and then for each of the eight specific dimension-level scores as the outcomes of interest. The presentation of the models below uses the generic $Score_{ti}$ outcome of interest for simplicity's sake, but the Results chapter discusses results for varying outcomes of interest.

4.1.1 Determining the best time function. A key decision in estimating growth is the most appropriate way to model time. The data in the MET project spans two years, so a first question is if the models should include separate intercepts and slopes for each of the two years or if there should only be one slope and one intercept across years in the MET project. It is reasonable to think that a school district might want to have a separate growth trajectory for each year. For example, the district may want to use growth as one of multiple measures of teacher effectiveness for teacher evaluations. However, a concern with this approach is the ability to measure growth reliably with the data for only one year. As such, the first step is to run a single HLM analysis that includes only the $week_{ti}$ variable at level 1. The dependent variable is mean FFT score within occasion. Table 4.1 provides the specifications for the sub-models in this preliminary analysis.

Table 4.1 Reliability Growth Model Specifications

Model	Specification
Level 1	$Score_{ti} = \pi_{0i} + \pi_{1i}(week_{ti}) + e_{ti}$, where $e_{ti} \sim N(0, \sigma^2)$ (4.1a)
Level 2	$\pi_{0i} = \beta_{00} + r_{0i}$, where $r_{0i} \sim N(0, \tau_{00})$ (4.1b)
	$\pi_{1i} = \beta_{10} + r_{1i}$, where $r_{1i} \sim N(0, \tau_{11})$ (4.1c)
Composite	$Score_{ti} = \beta_{00} + \beta_{10}(week_{ti}) + r_{0i} + r_{1i}(week_{ti}) + e_{ti}$ (4.1d)

I run the model specified in Table 4.1 on three different data sets: 1) all occasions in the first year of the MET project, 2) all occasions in the second year of the MET project, and 3) all occasions across both years of the MET project. The results of interest from this initial analysis are the reliabilities for the design-specific growth parameter estimates in each subset of the data. If it is the case that the reliabilities for the single year growth parameter estimates are relatively

close to the across years estimate, then it makes sense to model growth separately for each year in the MET project as districts may find this level of growth analysis useful. However, if modeling growth across years yields growth parameter estimates with much higher levels of reliability, then a better choice is a single growth trajectory over the course of the entire MET project[15]. The Results chapter provides the details of this preliminary analysis as an entry point for the full discussion of reliability. However, for proper interpretation of the remainder of this chapter and the HLM results in Chapter 5, it is important for the reader to know that the most reliable and best model fit is the result of a linear time function that models growth across both years of the MET project data.

4.1.2 The unconditional means model. The analysis begins with an unconditional means model. This model provides information about mean FFT scores, whether teachers vary in these scores generally, and whether it seems reasonable to think that teachers vary in these scores over time. If the variability in observation scores across people is not statistically significant, there is little point in trying to explain the variability with a time trend. Table 4.2 lists Equations 4.2a-4.2c. In these equations, *t* denotes time, or occasion, and *i* references individuals, or teachers. As such, the outcome of interest in the level-1 model, $Score_{ti}$, represents the score at time *t* for teacher *i*[16].

---

[15] Another important consideration is modeling time in a linear fashion or something more complicated such as a quadratic function. I used F-statistics to compare both linear and quadratic functions of time. Of the 458 teachers in the sample, the quadratic function was only statistically significantly better than the linear model in less than 1% of cases across years. A similar analysis within each year of the data yielded similar results. Specifically, 3% and <1% of cases fit the quadratic model better than the linear with the year 1 and year 2 data respectively. Thus, this study includes a linear function to model time.

[16] It should be noted that a three-level HLM analysis was attempted to model dimension-level variability. However, only a small amount of variation in FFT scores is explained by time. As a result, including the parameters

Table 4.2 Unconditional Means Model Specifications

Model	Specification
Level 1	$Score_{ti} = \pi_{0i} + e_{ti}$, where $e_{ti} \sim N(0, \sigma^2)$ (4.2a)
Level 2	$\pi_{0i} = \beta_{00} + r_{0i}$, where $r_{0i} \sim N(0, \tau_{00})$ (4.2b)
Composite	$Score_{ti} = \beta_{00} + r_{0i} + e_{ti}$ (4.2c)

Equations 4.2a-4.2b provide the models for each of the specified levels in the HLM, and the composite model (4.2c) includes the substitution indicated in the first two models and presents all the relevant parameters of interest in one equation. Each of these models consists of two parts, the structural and the stochastic portions of the model. The stochastic terms, also referred to as random effects, are those assumed to vary. The structural terms are the fixed effects. For example, β_{00} makes up the fixed effects portion of Equation 4.2c and the remaining two terms define the random effects portion of the equation.

Next is a more careful narrative of the meaning behind the terms in Table 4.2, and a visual interpretation in Figures 4.1 – 4.2 of a simplified case.

- π_{0i} is the expected mean FFT score for an individual teacher *i*,
- β_{00} is the expected grand mean of FFT scores across all occasions and teachers,
- r_{0i} is the deviation of teacher *i* from β_{00} (i.e., a between-teacher random effect), and
- e_{ti} is the deviation of occasion *t* from teacher *i's* mean FFT scores (i.e., a within-teacher random effect).

necessary to investigate dimension-level variance as well as allowing for those variance components to be random, was simply too demanding on the model. In other words, the model specification asked HLM7 to model variability where there was not sufficient data to model such variability. As a result, estimating two-level models with each of the different outcome variables occurs later in this chapter to investigate dimension-specific variability instead.

In addition to the parameters of interest from the composite model are the related variance components from the random effects within the HLM.

- τ_{00} is the between-teacher variance or the scatter of teachers' expected FFT scores around the grand mean of FFT scores, and
- σ^2 is the within-teacher variance, or the distribution of each teacher's FFT occasion-specific scores around his or her own expected FFT score.

Figure 4.1 provides a simplified illustration for better interpretation of these parameters. In it, I consider a subset of only three distinct teachers (i=1, 2, and 3). The three orange lines represent the mean FFT score across occasions for each example teacher. The green line represents β_{00}, or the expectation of the grand mean for all three teachers. Each orange line (π_{0i}) depicts a teacher-specific expected FFT score across occasions, and the space between each orange line and the green line is the teacher-specific deviation (r_{0i}). The term τ_{00} (not pictured), characterizes the variability of the r_{0i} terms across teachers. Notice that although occasions are technically individual points in time, change related to time is not modeled in the unconditional means model. Instead, each person's trajectory is a flat line, and variation across persons occur around β_{00} only.

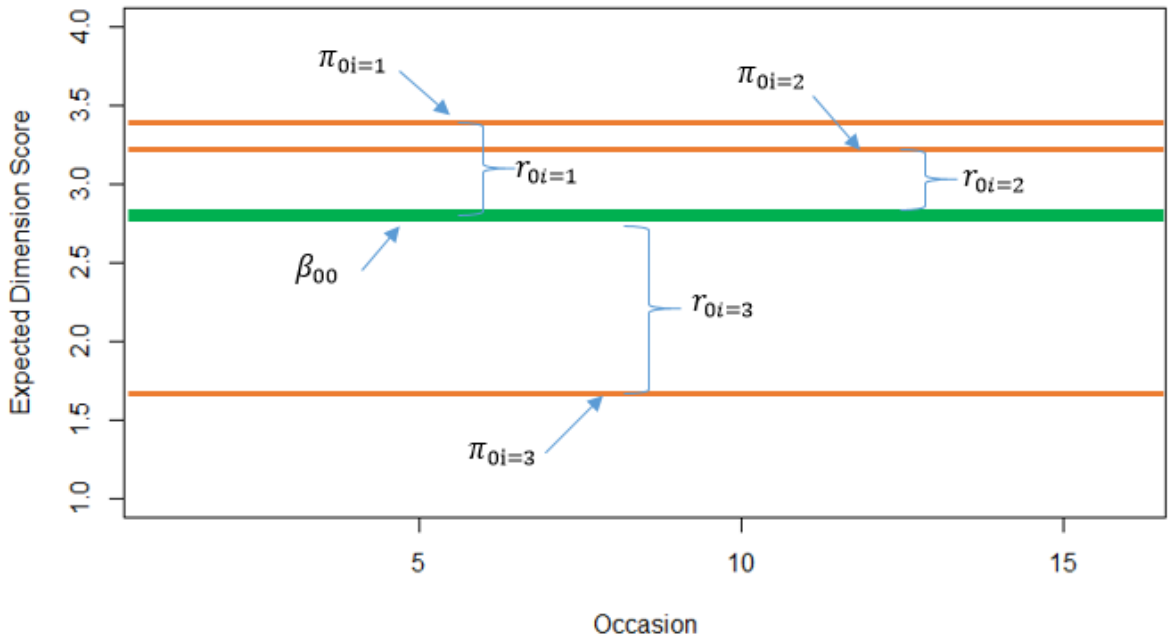


Figure 4.1 Illustration of level-2 unconditional means parameters

Figure 4.2 includes the observed mean FFT scores for six specific occasions (the red dots) so that this figure includes both the level-1 and the level-2 parameters for a single teacher (i = 3). The distance between each red dot and the bottom orange line ($\pi_{0i=3}$) is an estimated individual error term for that specific teacher and occasion (e_{ti}). The term σ^2 (not pictured) characterizes the variance of these distances. These deviations (or variability within teacher over time) are, in part, caused by measurement error. Importantly, salient changes in teacher practices over time also influence this deviation.

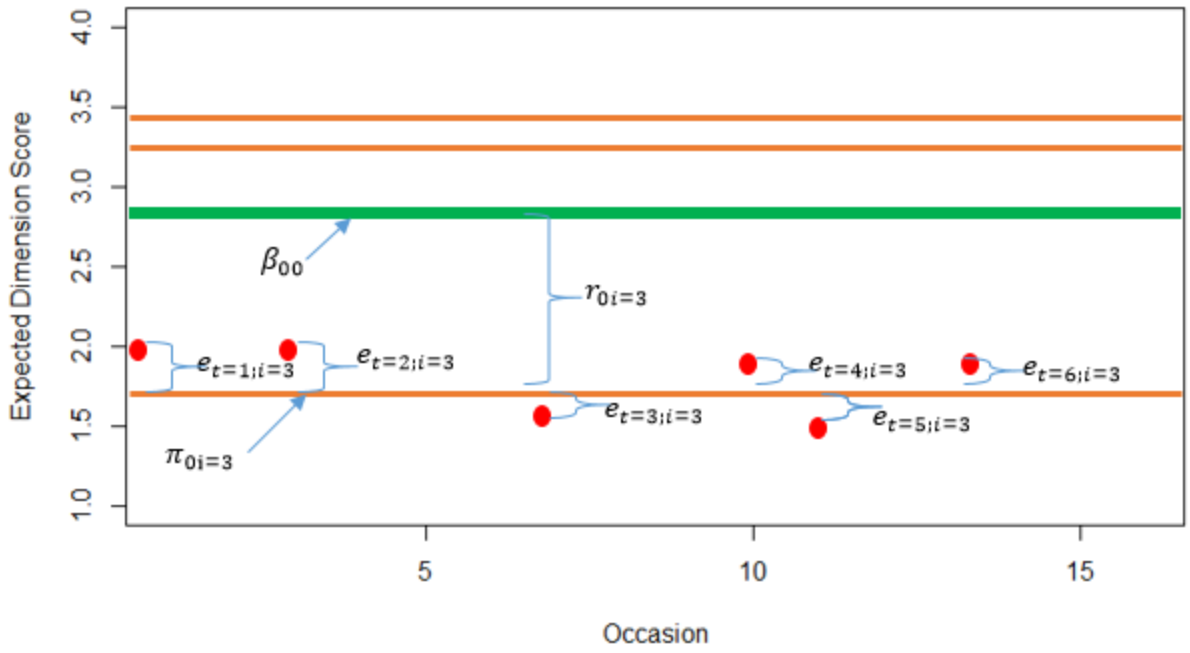


Figure 4.2 Illustration of level-1 and 2 unconditional means parameters

Following estimation of the unconditional means model, it is important to consider if the variance components are statistically significant. If the level-2 variance component (τ_{00}) is insignificant, there is no need to apply HLM. However, if all of the variance components are determined to be significant, there is evidence to suggest that the clustered nature of the data has an influence on the outcome of interest. A chi-squared test statistic assesses the statistical significance of the level-2 variance component.

In addition to testing the statistical significance of the variance components, it is also important to calculate the proportion of variance in FFT scores explained at each level of the unconditional means model. These calculations occur with Equations 4.3a-4.3b.

$$\frac{\sigma^2}{\sigma^2+\tau_{00}} = \text{the proportion of variance within teacher; and} \quad (4.3a)$$

$$\frac{\tau_{00}}{\sigma^2+\tau_{00}} = \text{the proportion of variance between teachers} \quad (4.3b)$$

The unconditional means model informs whether teachers vary around mean FFT scores and how much of that variance occurs within teachers and between occasions. From here, we transition to an investigation of change over time.

4.1.3 The unconditional growth model. The unconditional growth model builds on the unconditional means model by adding a time function at level 1. This is done to answer the following questions: What is the estimated average linear growth trajectory of change in FFT scores for teachers for each year in the MET project (i.e., β_{00} and β_{10})? In addition, do teachers vary in their rate of growth in FFT (i.e., the diagonal elements of the Tau matrix)? Table 4.3 lists the equations for the unconditional growth model in Equations 4.4a-4.4d.

Table 4.3 Unconditional Growth Model Specifications

Model	Specification
Level 1	$Score_{ti} = \pi_{0i} + \pi_{1i}(week_{ti}) + e_{ti}; e_{ti} \sim N(0, \sigma^2)$ (4.4a)
Level 2	$\pi_{0i} = \beta_{00} + r_{0i}; r_{0i} \sim N(0, \tau_{00})$ (4.4b) $\pi_{1i} = \beta_{10} + r_{1i}; r_{1i} \sim N(0, \tau_{11})$ (4.4c)
Composite	$Score_{ti} = \beta_{00} + \beta_{10}(week_{ti}) + r_{0i} + r_{1i}(week_{ti}) + e_{ti}$ (4.4d)

The new predictor variable in the model characterizes time as a single linear predictor for week. The unconditional growth model allows for teachers to vary not only on their *level* of mean FFT score in the MET project, but also on their *rates of change*. Brief explanations of the level-2 parameters and the related variance components appear below:

- β_{00} is the expected grand mean of FFT scores at the beginning of the MET project across all teachers,

- $\beta_{10}$is the expected mean growth rate across teachers during the project,
- r_{0i}, and r_{1i} are the teacher-level random effects for the prior two parameters[17].
- τ_{00} is the between-teacher variance in expected FFT score at the beginning of the MET project,
- τ_{11} is the between-teacher variance in expected rate of change, and
- σ^2 is the within-teacher variance, or the distribution of each teacher's FFT occasion-specific scores around his or her own FFT score trajectory.

The variance components in the unconditional growth model separate variation due to within-teacher factors over time (level-1 variation) from variation due to between-teacher factors (level-2 variation). The level-2 variances now summarize between-person variability in FFT scores at the beginning of the MET project and the rate of change during the project. If there is significant variation around the mean growth parameters, we might next explore *why* some teachers exhibit steeper growth than others. In reconsidering Figure 4.2, the new level-2 variance components now allow the orange and green lines to be sloped rather than flat as they appear in the illustration for the unconditional means models.

Next is an examination of the amount of level-1 variation in FFT scores accounted for with the $week_{ti}$ variable. It is hypothesized that $week_{ti}$ accounts for some of the level-1 variance identified in the unconditional means model. As a result, the level-1 variance component of the unconditional growth model now provides an estimate of the residual variance

[17] I used deviance statistics to compare this model that allows both of the level-2 effects to vary randomly to two other models: one that only allowed the intercept to vary randomly and one that only allowed the slope to vary randomly. In each of these comparisons, the p-value was <0.001 indicating that the model with both random effects was the best fit to the data.

at level-1. In other words, comparing the level-1 variance component from the unconditional growth model to the unconditional means model provides information regarding how much $week_{ti}$ accounts for level-1 variance. This calculation is performed using Equation 4.5.

$$\frac{\sigma^2_{(\text{Eq4.2a})} - \sigma^2_{(\text{Eq4.3a})}}{\sigma^2_{(\text{Eq4.2a})}} \quad (4.5)$$

The results form Equation 4.5 provide a sense of whether $Week_{ti}$ explains much of the within-teacher variability in FFT scores. From here, it is important to examine whether the level-2 variance components are statistically significant. The null hypothesis here is that each of the level-2 variance components are non-significant, or that the null (unconditional means) model is the correct specification. The alternative hypothesis is that one or more of these variance components are statistically significant and explain some of the variability in FFT scores or that the unconditional growth model is a better way to specify the model. The chi-square statistic tests the statistical significance of each of these components. If the residual level-2 variance terms are significant according to the chi-square test, the next step is to include level-2 predictors that might explain heterogeneity in the level-one parameters.

4.1.4 The novice model. The unconditional growth model allows for the estimation of longitudinal growth trajectories across individual teachers. Recall that the unconditional growth model provided an answer to the following question: What is the estimated average longitudinal growth trajectory of change in FFT scores for teachers in the MET project, and do teachers vary in terms of their growth? We now turn to exploring that variability across teachers. The only difference between the unconditional growth model and the novice model is the introduction of a

time-invariant, level-two predictor variable, $Novice_i$. This variable now appears in each of the level-2 equations (4.4b-c). The purpose of this model is to answer the following question: On average, do novice teachers exhibit different growth trajectories than more experienced teachers?

Since the only difference between the unconditional growth model and the novice model is that change over time is now modeled conditionally on a teacher's $Novice_i$ status, there are only two additional parameters to explain in this model, plus a new interpretation of the Level-2 error terms.

- β_{01} is the difference between novice and experienced teachers in expected FFT score at the beginning of the MET project,
- β_{11} is the difference between novice and experienced in expected growth rate,
- r_{0i} and r_{1i} are now the teacher-level *residual* variance (after taking into account whether a teacher is novice or not) in the prior two parameters.

The two coefficients on the $Novice_i$ predictor capture whether growth trajectories differ for novice and experienced teachers. For example, we might expect a new teacher to grow at a faster rate during the MET project than a more seasoned teacher. As before, a chi-square test examines whether the level-two variances are statistically significant. It is unlikely that the $Novice_i$ variable alone will explain all of the across-teacher variability, and so the expectation is that the significance of these variances will not change much from the unconditional growth model.

The results from the novice model provides evidence regarding the potentially steeper growth of novice teachers suggested in the conceptual framework. However, all of the analysis up to this point used average score across dimensions as the outcome of interest. Prior research specifically calls out particular behaviors as those for which novice teachers are most expected to

vary from more experienced teachers. As such, the HLM analysis is repeated with each of the FFT dimensions as the outcomes of interest to investigate potentially different rates of growth for novice teachers on managing classroom procedures and managing student behavior.

4.1.5 Gain scores. The analysis up until this point assumes that teachers grow in their practices both within and across years. As such, the estimated growth trajectories use each occasion in the MET project data set as individual time points that provide information about teacher practices. However, as discussed in the literature review, single observations with single raters typically have relatively low levels of reliability (Kane & Staiger, 2012). Much prior research with observation protocols averages scores across multiple occasions not only in order to achieve more reliable measures of teacher practices but also because it is assumed that any change in observation scores across occasions is due to measurement error rather than salient changes in teacher practices. Before moving on to an investigation into the factors that influence the reliability of growth parameter estimates, I first conduct one final HLM analysis.

In this approach, I assume that salient change in teacher practices occur across years, and that any changes in FFT scores observed across occasions within each year is due to measurement error. As such, I average all of the FFT scores within each year of the MET project and run an HLM analysis with a slightly modified novice model. It might be the case that estimating growth in this way yields more reliable estimates of growth as the basis for the growth estimates are more reliable themselves.

In the revised approach, the outcome of interest is mean FFT score across all occasions within a year, so each teacher has two scores in the dataset. Since this approach requires that teachers have observation scores in each year of the MET project, the available decreases from

N = 458 to N=441. Rather than modeling time with the $week_{ti}$ variable from the previous models, time is now modeled with a dummy variable for the second year of the MET project ($Y2_{ti}$). The interpretation of β_{00} in this model is the expected mean FFT score for Year 1 of the MET project, and the interpretation of β_{10} is the change in mean FFT score between the two years of the MET project. In other words, β_{10} is still a growth parameter estimate, but now it is an estimate for the gain score across the two years. As before, the novice model simply adds a variable to control for being a novice teacher to each of the level-2 equations, and the estimates of the related parameters (β_{01}, β_{11}) indicate if there is a difference in Year 1 scores or change between years for novice as opposed to experienced teachers.

Those in charge of designing observation systems must make a choice regarding the use of additional observation occasions. If it is assumed that teachers make meaningful changes in their practices both across and within school years, then estimates of growth will be based on less reliable measures, but more of them. In contrast, if it is assumed that growth only occurs within year, then growth estimates are based on more reliable measures, but fewer of them. This final HLM analysis allows for comparison between the two approaches. It may be the case that one is better than another in identifying differences in rates of growth between groups of teachers. If that is true, it is important information for those designing observation systems.

4.1.6 Growth estimate validity check. A unique element of the MET project dataset is the fact that the recorded lessons as well as the MET rater observation scores are available to researchers. As such, the present analysis includes a brief validity check into the HLM and gain score results. I performed an independent analysis of the videos for the four teachers with the largest magnitude of growth parameter estimates in both positive and negative directions. This

means that I viewed the 68 videos submitted by the eight teachers in the MET project who supposedly changed the most in either the positive or negative direction over the course of the two years in the project.

In this process, I blinded myself to the estimated slope parameter for each teacher, and scored each teacher's videos with the FFT in order of submission. Although I did not go through any formal training regarding the use of the FFT, I used the protocol materials provided to MET raters to assign scores on all eight dimensions to each video for each teacher. Just as described in the MET project User's Guide (Bill & Melinda Gates Foundation, 2013), I watched the first fifteen minutes of each video and then skipped to watch minutes 25 – 35. There were two exceptions to this guideline. First, there were three instances when the video recordings did not exceed 35 minutes. In these instances, I watched the first 15 minutes and then watched from minute 25 to the end of the video. The User's Guide provided no information regarding what raters did in these scenarios, so I can only assume that my process was the same as theirs. Second, the server hosting the videos crashed frequently while I watched these videos. As a result, it was not uncommon for me to miss a few seconds when trying to start the video where I had last left off multiple times a video. I assume that this process did not affect the original MET raters as this was highly problematic for my scoring process.

After scoring each video, I recorded the mean score for each occasion and then a mean score for each year. The occasion data was used in eight, teacher-specific ordinary least squares (OLS) linear regressions in which mean FFT scores were regressed on week. I then compare the estimated slope values from the HLM to those from the teacher-specific OLS regressions. Additionally, I compared my gain scores to the gain scores from MET raters. Finally, I provide commentary on the level of agreement between my scores and those in the MET project dataset.

Following presentation of these results, I turn to an investigation of the reliability of the growth estimates and the observation system designs that yield the most reliable information about growth.

4.2 Reliability of Growth Parameters

The reliability of the growth parameter estimates in this study is important for two key reasons. First, if any school or district wanted to use growth estimates for any sort of decision-making dependent on rank order of growth, it is important that those statistics be able to make reliable distinctions among individual teachers. More than that, it is also important for research purposes that the parameters be reliable enough to detect if differences in growth actually exist among groups of teachers—particularly if the differences are relatively small. Of interest in the current study are the differences between novice as opposed to experienced teachers.

Reliability indices provide information about how a measurement is expected to vary across replications of a measurement procedure. In other words, reliability estimates attempt to quantify the precision of a measurement (Haertel, 2006). Upon repeated measures, we wonder how sure we are to get the same estimate and how consistently we can rank order teachers. Conceptually, reliability indicates the proportion of variance in an outcome of interest that is due to true change in the "signal" rather than "noise". More specific to the current context, it indicates how much of the change in FFT scores is due to true change in teacher practices (signal) as opposed to measurement error (noise). Recall that there are two ways to consider reliability in this dissertation: reliability of FFT score status on a single occasion and reliability

of the rates of change in FFT scores over occasions. It is important to consider the manifestations of measurement error for both of these ways of thinking about reliability.

A key source of measurement error for status scores is that from raters. A rater might pay more or less attention to any given video, thus influencing the amount of measurement error in the individual observation scores. This source of measurement error for status FFT scores can be mitigated by adding an additional rater to each occasion. Another source of error in the current project is the fact that lessons are nested within teachers. The MET project allowed for teachers to select the lessons they recorded. Since multiple teachers were not observed teaching the same lessons, there is no way to disentangle the effects of true teacher behavior from idiosyncrasies related to particular lessons.

When considering measurement error as it applies to growth parameter estimates, part of the variability in slopes is due to choice of lesson on a given occasion. In other words, lessons are not only nested within teachers, but also nested within occasions. Consider the day before winter break as opposed to a typical Wednesday in February. Some variance in scores between these two occasions may be due to true teacher change, but some will also likely be due to the specific occasion on which a lesson was delivered. As such, an idealized measurement design would include a set of common lessons that all teachers would deliver on multiple different days. This might be accomplished with lessons focused on topics such as mathematical reasoning (e.g. modeling with mathematics, constructing viable arguments and critiquing the arguments of others) or reading comprehension. Unfortunately, this is not the design of the MET project.

In the current context, variability due to lesson cannot be disentangled from variability due to teacher. Additionally, variability due to lesson cannot be disentangled from variability due to occasion. As such, all of this noise is included in the estimated signal in the current project. If

the ratio of signal to noise is poor, then we feel less confident in our ability to attain similar estimates upon repeated measures. This is important if, for example, a district wanted to use teacher growth rates as a way to identify teachers who should receive bonuses. If it were the case that growth rates were not measured very reliably, then the district would not have much confidence in their ability to correctly identify the most deserving teachers. Additionally, the more measurement error, the less sure we are about if growth differs for various groups of teachers. For example, if the difference in the growth rates for novice versus experienced teachers is very small, they will be undetectable if the growth parameter estimates have low reliability.

Before discussing the details regarding how the reliability of the growth parameter is calculated, it is important to more carefully consider reliability in this specific context. Recall that the MET data is time-unstructured and unbalanced. As a result, any teacher with a unique combination of number of occasions and spacing of those occasions can be conceptualized as representing a unique design for measuring teacher practices. The overall reliability of the growth parameter from the HLM gives the average reliability across all of these measurement designs. The discussion of reliability presented in the Results chapter includes the overall reliability as well as a description of the full distribution of reliabilities from the different measurement designs present in the MET project.

The reliability (ρ) of an estimated growth parameter, here indicated by θ to provide a general case, is given by Willet (1988) as follows:

$$\rho(\theta) = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \left[\frac{\sigma_\varepsilon^2}{SST}\right]} \qquad (4.6)$$

As discussed earlier, reliability is generally thought of as the proportion of signal to observed variability. Thus, σ_θ^2 is the estimate of signal, and $\left[\frac{\sigma_\varepsilon^2}{SST}\right]$ is the estimate for noise. Each of the terms in the above formula can be further explained as such:

- σ_θ^2 is the true variance of teacher-specific growth parameters. In other words, it is the signal, or true change in teacher behaviors. Each measurement design has a unique growth parameter estimated with weighted least squares in HLM. In using weighted least squares estimation, each slope estimate is weighted proportional to its precision. This weighting takes into account the fact that teachers vary in both the number and spacing of their occasions. A population-level estimate of growth parameter variance is taken across teachers as an average of these weighted values.
- σ_ε^2 is the variance of the level-1 errors, e.g. how far each teacher's recorded observation scores deviate from the expected score for that particular design. For any given teacher and occasion, there is deviation between the observed score and the expected score. The mean of these deviations is the teacher-specific mean-squared error. Similar to the variance of the growth parameter, σ_ε^2 represents the average of the mean-squared errors across all teachers.
- SST is the sum of the squared deviations of observation occasions about their mean $[\sum_{i=1}^{t}(t_i - \bar{t})^2]$. It is the quantification of the number and spacing of observations for each unique measurement design. This value moderates the effect of measurement error on the reliability estimate. In other words, reliability and SST are positively related; as SST increases, so do the number of occasions and the spacing between occasions, on average. This, in turn, increases reliability.

Before describing the considerations of reliability in the context of the current study, the next section provides an explanation of the last term in the denominator.

4.2.1 Derivation of the formula for reliability of a growth parameter. The growth parameter in the HLM specifications provided earlier in this chapter are level-1 slope parameters. Due to this specification, we can think of the growth parameter as a slope parameter in a simple linear regression. Thus, it is important to understand the least squares estimator of the slope parameter to fully appreciate the calculation of the reliability of a growth parameter. Consider a simple linear regression specified as follows:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ (4.7).}$$

In OLS regression, the estimator for β_1 is expressed in Equation 4.8, and the sampling variance of. β_1 is defined by Equation 4.9.

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \text{ (4.8)}$$

$$\text{Var}(\beta_1|x) = \frac{\sigma_\epsilon^2}{\sum(x_i-\bar{x})^2} \text{ (4.9)}$$

We can now see that the variance of the least squares estimator of the slope parameter is equal to the quantity we see in the second term in the denominator of Equation 4.6 when x_i is time and $\sum(x_i - \bar{x})^2$ is expressed as $\sum(t_i - \bar{t})^2$ referred to as SST. Willett (1998) discusses SST in more detail to explain this concept.

> The quantity SST is a function of *both* the number of waves of data collected in the growth study *and* the relative spacing of those waves. Thus, the reliability of the growth measurement can be improved *either* by collecting more waves of data *or* by judiciously altering the spacing between the waves, or both. Moreover, because SST is a *quadratic* function of the observation times, it increases very rapidly as the occasions of measurement are manipulated. Therefore, the manipulation of SST is a much more effective and convenient method of increasing the growth reliability in practice. (pp. 404-405).

SST is an integral part of the formula for the reliability of a growth parameter because it is a factor in collecting teacher observation scores that can most easily be controlled in the local context. Maximizing SST is a way that stakeholders can increase the reliability of their growth estimates and thus an important consideration when designing a teacher observation system. The final sections of this chapter consider how reliability might vary in the current data set as well as other hypothetical contexts in which the reliability of growth in teacher observation scores might differ from what is in the current project.

4.2.2 Understanding reliability in the current context. The design of the MET project included four occasions per teacher per year for a total of eight occasions per subject spread across two school years. The preliminary reliability analysis not only informs the models used in this project but also provides an entry point into understanding the ways the number and spacing of occasions influences the reliability of growth parameter estimates. Recall the following information about the spread of occasions in the MET project:

- Year 1: four occasions spread over nineteen weeks (i.e. from February to June),
- Year 2: four occasions spread over thirty-one weeks (i.e., from October to June) and
- Both years: eight occasion spread over fifty weeks.

These reliability estimates provide a first step in understanding how the spacing and number of occasions might influence reliability estimates in the current project. The contrast from the first

to the second estimate of reliability highlights what can happen to reliability when the number of occasions remains the same but the spacing gets wider. The contrast from the first to the third indicates how both the number and spacing of occasions affects reliability. Beyond discussion of this cursory information, this section also provides more detailed investigation of SST, the number and spacing of observations, the relationship between SST and the number and spacing of observations, and how all of these factors relate to reliability of the growth parameter.

4.2.3 A best case for reliable growth parameters. We can leverage the distribution of measurement designs in the MET project to gain an understanding for contexts that yield growth parameters with higher reliabilities. This section describes, in detail, the measurement designs with the highest levels of reliability for the growth parameter estimates in the MET project as well as conducts a thought experiment regarding increasing reliability by adding an additional rater to each occasion.

Recall that reliability of the growth parameter is a function of three components: the variability in the growth parameter itself, variance of the level-1 error term, and SST. The current data have little variability in growth parameter estimates. District or school leaders cannot influence this variability in a local context. Teacher practices either change or not. Thus, this investigation of potential reliabilities does not include consideration of a change in this variance. However, it is important to remember that the variance of the growth parameter will vary by context.

School and district leaders do have some power to influence variance in the error term as well as the number and spacing of occasions to some extent. The MET project is likely a best-case scenario with respect to the number of observations, as four lessons per teacher per subject

per year for all teachers were included in the study design. This number of observations is potentially unrealistic in many local contexts, but the presence of such rich data in the MET project allows for investigation of what levels of reliability could be reached should a local education agency decide to devote enough resources to collect this many annual observations for all teachers or adopt, for example, a three-year review cycle that enabled deeper analysis for one-third of teachers annually. Thus, this final analysis characterizes the SSTs in the top quartile of the distribution of SSTs and discusses the related reliabilities before delving into a thought experiment regarding adding an additional rater to each occasion and the potential effects that an additional rater might indirectly have on the reliability of the growth parameter estimate.

The Results chapter shows that the ratio of signal to noise in growth estimates across teachers in the MET project is rather poor. As a result, the reliability of the growth parameters is necessarily low. These explorations of the ways that SST and measurement error might influence reliability provides a better context for understanding the potential for the reliability of a growth parameter estimate that could be attained using the methodologies presented in this dissertation. This not only provides helpful information for those designing observation systems but also gives a better context for understanding the results of the current project.

4.2.4 Reliability of a gain score. In this last section, I discuss the reliability of gain scores as opposed to growth parameter estimates. If one were to decide that growth is only measureable across years after averaging across multiple within-year observations, then it is important to know how the reliability of such gain scores compares to those of growth parameter estimates. Willet (1989) provides the following formula to define the reliability of a difference (D), or gain score:

$$\rho(D) = \frac{\sigma_{x_1}^2 \rho(x_1) + \sigma_{x_2}^2 \rho(x_2) - 2\sigma_{x_1}\sigma_{x_2}\rho x_1 x_2}{\sigma_{x_1}^2 + \sigma_{x_2}^2 - 2\sigma_{x_1}\sigma_{x_2}\rho x_1 x_2} \qquad (4.10).$$

The terms in Equation 4.10 are defined as follows:

- $\sigma_{x_1}^2, \sigma_{x_2}^2$ are the variances of the scores at time 1 and time 2 respectively,
- $\sigma_{x_1}, \sigma_{x_2}$ are the standard deviations of the scores at time 1 and time 2 respectively,
- $\rho x_1 x_2$ is the correlation between the time 1 and time 2 scores, and
- $\rho(x_1), \rho(x_2)$ are the reliabilities of the time 1 and time 2 scores.

Notice that similar to the formula for the reliability of a growth parameter estimate, two factors that drive this reliability formula are variation in observations scores and reliability/measurement error of status scores. If there is low variance in observation scores and low reliability of status scores, the reliability cannot be very high. However, if there is notable difference in the levels of change among individuals, gain scores can be quite reliable (Rogosa et al., 1982; Rogosa & Willett, 1983). Similar to the previous analysis, I not only provide information about the reliability of the gain scores in the current analysis, but also an explication of the best case for the reliability of a gain score if additional raters were employed to improve the reliability of status FFT scores.

4.3 Conclusion

The methods described in this chapter outline an approach for identifying change in teacher observation scores as measured by the Danielson Framework for Teaching. In addition, information regarding detecting if that change is systematically different for novice teachers as

opposed to their more experienced counterparts is included. The investigation begins generally before moving on to a more specific analysis that provides information regarding how that change might differ for the particular dimensions of the FFT. Finally, a gain score analysis illustrates another approach to measuring change in teacher practices over time. The chapter closes with a discussion of the various considerations of the reliability of growth parameters in the current dataset. This exploration provides a context in which to interpret the results of the current study as well as some guidance for those designing their own classroom observation systems. The following chapter provides the results from applying these methods to the data in the MET project.

5.0 Results

The results and relevant discussion presented in this chapter answer the previously defined three research questions in this dissertation. The results related to the first two questions appear in Section 5.1 with a presentation of the estimates from the unconditional means, unconditional growth, and novice models with mean FFT score as the outcome of interest. Next is a presentation of the results from the same models with dimension-specific scores as the outcome of interest in Section 5.2 and a continued discussion of the second research question in this dissertation. Sections 5.3 and 5.4 include the results from two small analyses investigating the validity of the results in the first two sections. In Section 5.3, a comparison of gain scores to HLM growth parameter estimates provides information regarding how the way one thinks about how growth occurs influences growth estimates. Section 5.4 includes an independent scoring analysis for eight teachers in the MET project as a validity check on the growth estimates presented in the chapter. Consideration of the third research question in this project occurs in Section 5.5 with analysis of the reliability of the estimated growth parameters in the MET project as well as the reliability under minimized error variance contexts. Finally, a brief conclusion to the chapter appears in Section 5.6.

5.1 Mean FFT Results

Table 5.1 presents the relevant output from HLM 7, the program in which these analyses are conducted, using restricted maximum likelihood estimation. The table includes the results for each of the three HLM models described in the previous chapter when mean FFT score across

dimensions on an occasion is the outcome of interest. This presentation allows for easy comparison of results across models. The key results from the unconditional means model appear in the first two columns of Table 5.1. The middle two columns include results from the unconditional growth model, and the final two columns show the same information for the novice model.

Table 5.1 Mean FFT HLM Results

	Unconditional Means Model		Unconditional Growth Model		Novice Model	
Fixed Effects	**Coefficient**	**SE**	**Coefficient**	**SE**	**Coefficient**	**SE**
Intercept (Mean FFT Score Beginning Y1)						
Experienced: β_{00}	2.49	0.01	2.49	0.02	2.53	0.02
Δ Novice: β_{01}	--	--	--	--	-0.24	0.05
Slope (Growth in FFT Score per week)						
Experienced: β_{10}	--	--	0.0001	0.0006	-0.0002	0.0007
Δ Novice: β_{11}	--	--	--	--	0.002	0.002
Random Effects	**Variance Component**	**p-value**	**Variance Component**	**p-value**	**Variance Component**	**p-value**
Level 1: e_{ti}	0.15	--	0.14	--	0.14	--
Level 2						
Intercept: r_{0i}	0.07	< 0.001	0.08	<0.001	0.07	<0.001
Slope: r_{1i}	--	--	0.00004	<0.001	0.00004	<0.001
Corr (β_{00}, β_{10})	--		-0.32		-0.31	
N	458					

5.1.1 Unconditional means model results. There is only one fixed effect in the unconditional means model, and it is the expected value of the mean FFT score across all occasions and teachers. The estimated value for β_{00} is 2.49 (p-value of < 0.001). Adding and subtracting 1.96 times the square root of τ_{00} provides a range of plausible values for the expected mean FFT score from 1.96 to 3.02. The range of plausible values suggests that for 95% of teachers, the expected mean FFT-dimension score on any given occasion is between two and three. This result is consistent with the illustrations of the data presented in Figures 3.2 and 3.3 in

the Data chapter. Next, and more interesting, is consideration of the variance components as they provide the most relevant information regarding variability in scores among teachers.

The variance components provide information about the proportion of variance in mean FFT scores at each level of the HLM. The statistical significance of these variance components indicate if HLM is necessary for the current analysis. Specifically, if the between-teacher variances are non-sigificant, then there is no variation in FFT scores attributable to the clustered nature of the data and OLS regression is sufficient.

The results in Table 5.1 indicate that the variance component at level-2 is statistically significant. Further, the results show that 68% of the variability in FFT scores occurs across occasions within teachers. Additionally, 32% of the total variance occurs between teachers. If roughly two-thirds of the variance in total FFT scores occurs between occasions and within teacher, it is possible that at least part of that variance is due to salient changes in teacher practices over the course of the MET project. This suggests that there is a significant amount of variance in mean FFT-dimension score that can be attributed to the clustering of data within teachers. As a result, it makes sense to move forward with further HLM analysis.

5.1.2 Unconditional growth model results. The unconditional growth model estimates an intercept that indicates the mean FFT score for the beginning of the MET project across all teachers (β_{00}), and a slope that indicates the mean rate of growth in FFT scores for all teachers across the full span of the MET project (β_{10}). The fixed effect estimates reported in Table 5.1 indicate that the fixed effect for mean FFT score at the beginning of the MET project is statistically significant, but the mean rate of growth is not. This means that, on average, there does not appear to be any growth distinguishable from zero in FFT scores over the course of the

MET project. However, just because the fixed effect indicates a mean growth rate of zero, does not actually mean that there is no change in observation scores for any individual teachers. It simply means that the average of the growth rates across teachers is not statistically different from zero. Thus, the variance components provide information regarding if there is any statistically significant variation in the rates of growth.

The variance components in this model suggest that there is statistically significant variation both in mean FFT scores at the beginning of the MET project as well as in the rate of growth, albeit in modest magnitudes. The range of plausible values for FFT scores at the beginning of the MET project ranges from 1.99 to 3.05. Similarly, the range of plausible values for the slope spans from -0.012 to 0.012. This means that 95% of teachers experienced growth rates between -0.012 to 0.012.

Proper interpretation of the plausible value range for the slope necessitates remembering that the unit of measurement for the growth parameter is weeks, and there are 50 weeks in the MET project. Thus, this plausible value range suggests that over the course of the MET project, we might expect the few teachers in the tails of the distribution to increase or decrease by as much as 0.6 points on average in mean FFT score. Recall from the Data chapter that the standard deviation in FFT scores is 0.47. Thus, it is possible for those teachers in the top or bottom 2.5% of the distribution of slope estimates to move up or down as much as 1.3 SD units in mean FFT score over the course of the MET project. However, such a large amount of growth requires the observations to span the full 50 weeks of the MET project, or for a teacher to submit videos in both the first and last weeks of the study. This is not the case for any teachers in the MET project.

In order to gain a better understanding of the amount of growth seen by these teachers in the extreme tails of the distribution, it is important to understand more about the distribution of estimated values for the slope parameter. Figure 5.1 shows a kernel density plot of the distribution of empirical Bayes estimates of the slope parameter. The mean of these slope parameter estimates is 0.000079 and the SD is 0.0028. Since these are shrunken Bayes estimates, the plausible value range is (-0.0054, 0.0056).

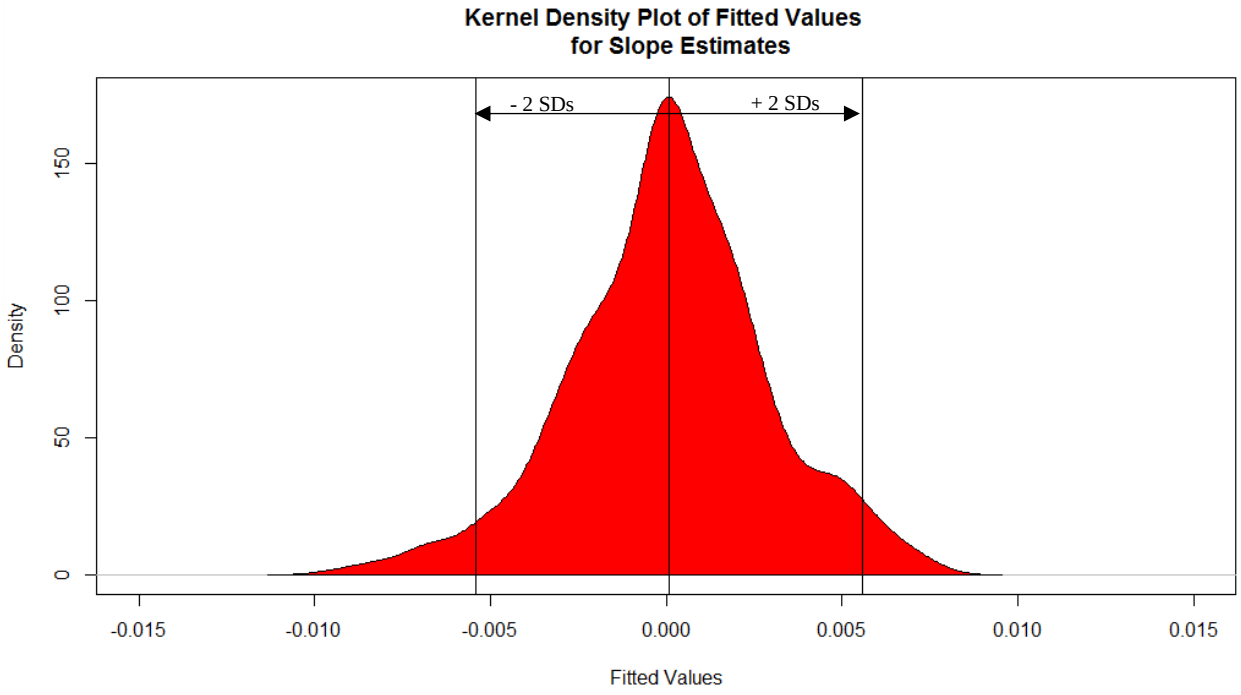


Figure 5.1 Distribution of empirical Bayes slope parameter estimates

The range for the top 2.5% of slope estimates runs from 0.0056 to 0.0075, and the range for the bottom 2.5% of slope estimates runs from -0.0094 to -0.0057. The teachers falling into these tails had an average of 8.5 (SD 2.1) observations in the MET project. The number of weeks spanned by these occasions runs from 27 to 41 weeks of the MET project, as opposed to the full 50. Further, the mean number of weeks spanned by the occasions of these teachers is 35.4. Thus, a more realistic characterization of the highest growth experienced in the MET project is 0.27

(0.0075*35.4) points in mean FFT score over the course of the MET project. This means that the teacher with the most extreme slope parameter estimates in the MET project experienced a change in FFT scores of about 0.57 SD units, on average over the course of the whole project.

Figure 5.2 illustrates the growth trajectories for a random sample of 50 teachers from the study. The figure shows that the vast majority of the intercepts fall between 2 and 3 and that most of the growth is close to zero, which is consistent with the results presented above. Also consistent with the results presented above is that there are some exceptions. The two cases in the random sample with the most extreme intercepts (1.89 and 3.03) are in blue. Similarly, the two cases with the most extreme slopes (-0.004, 0.007) are in pink. This figure illustrates what can happen with more extreme cases (those in pink and blue) as well as what is most expected (those cases in black).

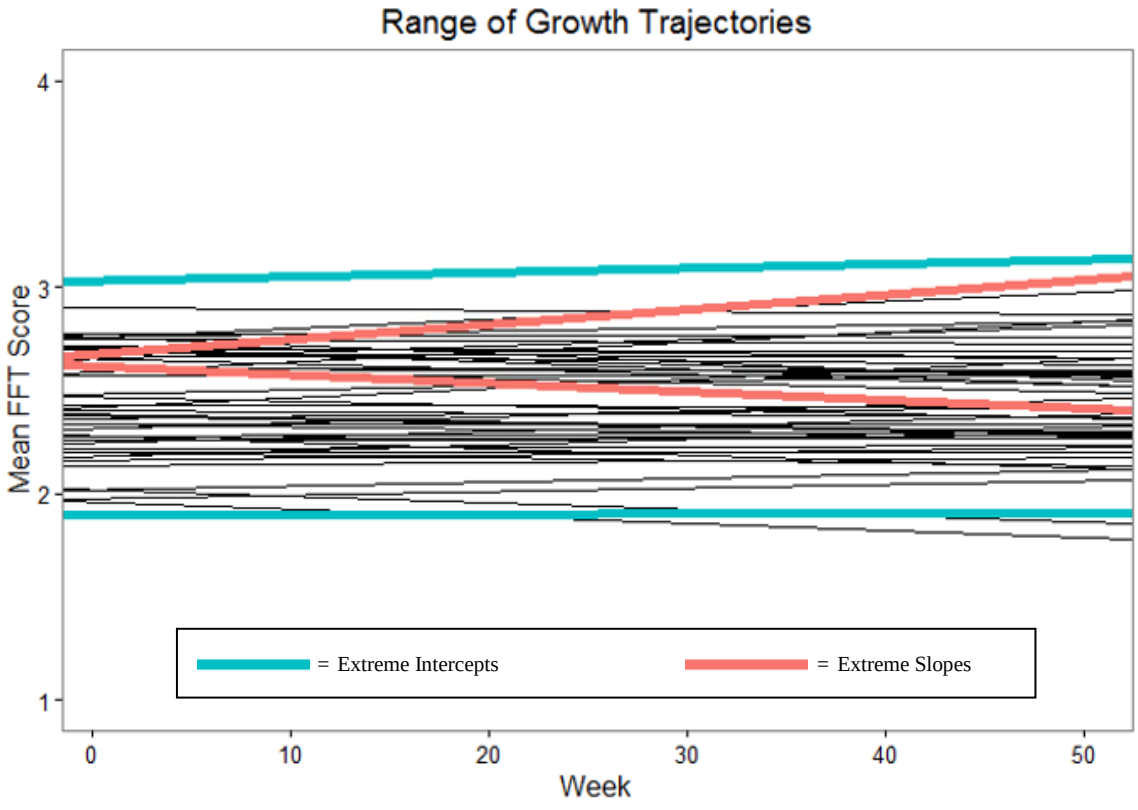


Figure 5.2 Illustrative growth trajectories

The discussion of the unconditional means model provided in the previous section indicated that about 68% of the variance in total FFT scores was within teacher (at the occasion level). The results from this model indicate that time accounts for about 7% of that 68%. In other words, the $week_{ti}$ variable in the unconditional growth model only explains about 5% of the variability in mean FFT scores. Thus, growth over time does not explain much of the within-teacher variability in FFT scores, but it does provide some explanation about the rates of change.

The correlation between the mean estimated intercept and slope is -0.32 in this model. This indicates that those teachers with higher initial FFT scores had lower rates of growth. This relationship is marginally lower in the next model, but the change of magnitude in the correlations is relatively small. This means that those teachers with higher beginning FFT scores have lower rates of growth, regardless of their status as novice or experienced teachers.

5.1.3 Novice model results. The final model in this sequence identifies if growth trajectories are different for novice and experienced teachers. The fixed effects estimates for the novice model reported in Table 5.1 indicate that we expect experienced teachers to receive mean FFT scores of about 2.53 at the beginning of the MET project, on average, and novice teachers to receive scores about one-quarter of a point (or about 0.5 SDs) lower. Similar to the unconditional growth model, the growth parameter estimates are non-significant both for experienced and novice teachers. This suggests that the level of FFT scores may differ between novice and experienced teachers, on average, but not the growth rates in this relatively short period of time.

Next, consider the variance components of the novice model. Both the slope and intercept have statistically significant variance components. As in the unconditional growth model, this

suggests that even though the average rate of growth is indistinguishable from zero, the significance of the variance component suggests that there is variability in the rates of change across teachers[18]. The magnitudes of the variance components between the unconditional growth and the novice model are similar, so we expect the same rates of growth in this model as the unconditional growth model.

These results do not provide empirical evidence to support the hypothesis that novice teachers grow at different rates than experienced teachers, but the work done up until this point only considers mean FFT scores within a given occasion. The conceptual framework suggests that novice teachers may be more variable in some dimensions of the FFT over others. It is possible that considering FFT scores in aggregate masks this variability, so the next section investigates potential differences in the results when the outcome of interest is the dimension-specific FFT scores.

5.2 Dimension-Specific Analysis

The previous results indicate that average growth in mean FFT scores is indistinguishable from zero, but the variability of that growth estimate indicates that a few teachers change about half of a standard deviation in FFT scores over the course of the MET project. Stopping here potentially masks patterns that might exist for some dimensions of the FFT. In order to investigate this point, and see if this dissertation can provide empirical evidence to support the

[18] Though one can estimate the proportion of variance in level-two outcomes explained by a set of level-two predictors, I opt not to do this because the level-two model simply consists of a single dummy variable and is not expected to account for a large portion of any level-two variability across teachers.

idea developed in the conceptual framework about the practices of novice teachers, the next section provides the results from the same sequence of analyses for each of the eight dimensions of the FFT. Table 5.2 repeats the two domains and related dimensions included in this study for easy reference. The dimensions in bold (managing classroom procedures and managing student behavior) are the two most closely related to those practices the conceptual framework highlights as most variable for novice teachers.

Table 5.2 FFT Dimensions and Domains in the MET Project

Domain	Dimension
Classroom Environment	• Creating an Environment of Respect and Rapport (CERR) • Establishing a Culture for Learning (ECL) • **Managing Classroom Procedures (MCP)** • **Managing Student Behavior (MSB)**
Instruction	• Communicating with Students (CS) • Using Questioning and Discussion Techniques (USDT) • Engaging Students in Learning (ESL) • Using Assessment in Instruction (UAI)

5.2.1 Dimension-specific unconditional means model. The unconditional means model provides information about if teachers vary in their dimension-specific scores, on average. Table 5.3 provides a summary of key results for all eight dimensions of the FFT. The Appendix D offers the full results from the HLM analysis.

Table 5.3 Unconditional Means Model Results—Dimension-Specific Analysis

	CERR	ECL	MCP	MSB	CS	USDT	ESL	UAI
Fixed Effects								
Coefficient: β_{00}	**2.67**	**2.46**	**2.64**	**2.74**	**2.59**	**2.18**	**2.39**	**2.24**
SE	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Variance Components								
Level 1: e_{ti}	0.28	0.32	0.29	0.24	0.29	0.36	0.32	0.36
Level 2: r_{0i}	**0.09**	**0.09**	**0.08**	**0.09**	**0.06**	**0.07**	**0.08**	**0.07**
Proportion of variance within teachers	0.76	0.78	0.89	0.73	0.83	0.84	0.80	0.84
Proportion of variance between teachers	0.24	0.22	0.22	0.27	0.17	0.16	0.20	0.16
N	458	458	458	458	458	458	458	458

*Statistically significant estimates ($p<0.05$) indicated in bold

As expected, every one of the dimension-specific scores is statistically significant as zero is not an option for scoring on the FFT. Further, the level-2 variance component for each dimension is all also statistically significant. This means that teachers do vary in their dimension scores for all eight dimensions. The proportion of variance within teachers and between teachers is relatively consistent across dimensions. Between 16% and 27% of the variability in dimension scores is between teachers, and from 73% - 89% of the variability is within teachers. Notably, MCP has one of the highest values for within-teacher variability. This is notable because this dimension is one predicted to have the most variability for novice teachers in the conceptual framework.

It is also important to note that since less than a quarter of the variance in the dimension-specific scores is attributed to the differences between teachers, the remaining three-quarters of the variance is due to differences within teachers. This is promising for further analysis that attempts to disentangle sources of variation in dimension-level FFT scores within teachers.

5.2.2 Dimension-specific unconditional growth model. Similar to the mean score analysis, the results from the dimension-specific unconditional means model suggest that there is significant variability in dimension-specific FFT scores at all levels of the HLM. Thus, the next analysis considers the unconditional growth model. Table 5.4 presents the key results for all eight dimensions. As before, the Appendix includes the full HLM results.

Table 5.4 Unconditional Growth Model Results—Dimension-Specific Analysis

	CERR	ECL	MCP	MSB	CS	USDT	ESL	UAI
Fixed Effects (SE)								
Intercept: β_{00}	**2.68 (0.02)**	**2.47 (0.02)**	**2.63 (0.02)**	**2.75 (0.02)**	**2.60 (0.02)**	**2.18 (0.02)**	**2.36 (0.02)**	**2.23 (0.02)**
Slope: β_{10}	-0.0004 (0.001)	-0.0002 (0.0008)	0.0008 (0.0007)	-0.0004 (0.008)	-0.0008 (0.008)	-0.0004 (0.0009)	0.0014 (0.0008)	0.0007 (0.0008)
Variance Components								
Level 1: e_{ti}	0.27	0.31	0.29	0.23	0.29	0.34	0.31	0.35
Level 2								
Intercept: r_{0i}	**0.10**	**0.09**	**0.08**	**0.14**	**0.07**	**0.07**	**0.08**	**0.09**
Slope: r_{1i}	**0.00003**	**0.00005**	<0.00001	**0.00007**	**0.00004**	**0.00005**	**0.00005**	0.00002
N	458	458	458	458	458	458	458	458

*Statistically significant estimates ($p<0.05$) indicated in bold

Similar to the analysis regarding mean FFT scores, the parameter estimate for scores at the beginning of the MET project are all statistically significant, and the growth parameters are all non-significant. Next, consider the variance components of the unconditional growth model. The results in Table 5.4 indicate that of the two dimensions the conceptual framework suggests will have the most variability, only one (MSB) has statistically significant variability among teachers.

The discussion of the unconditional means model provided in the previous section indicated that about one-quarter or less of the variance in dimension-specific scores was within teacher (at the occasion level). Now the level-1 model includes an additional control variable, the

associated variance component (e_{ti}) is a residual variance component. In order to understand how much of the level-1 variance is accounted for by the time variables included in the level-1 unconditional growth model, compare the residual level-1 variance to the total level-1 variance for all eight of the dimension-specific scores. Table 5.5 includes these results.

Table 5.5 Proportion of within Teacher Variance Explained by Unconditional Growth Model

	CERR	ECL	MCP	MSB	CS	USDT	ESL	UAI
Total Variance	0.28	0.32	0.29	0.24	0.29	0.36	0.32	0.36
Residual Variance	0.27	0.31	0.29	0.23	0.29	0.34	0.31	0.35
Proportion of variance explained	0.04	0.03	0.00	0.04	0.00	0.06	0.03	0.03

The results provided in Table 5.5 indicate that about 4% of the variance in dimension-specific FFT scores can be explained by the $week_{ti}$ variable in the unconditional growth model. This means of the 16% - 27% of variance at the occasion-level, about 4% can be explained by changes within teacher that are modeled by $week_{ti}$. In other words, from 0% - 1% of the variance in dimension-specific FFT scores is due to changes within teacher over time.

5.2.3 Dimension-specific novice model. The results from the unconditional growth model suggest that a modest amount of variance in dimension-specific FFT scores is due to differences at the end of year and growth for most of the dimensions. The final iteration of HLM analysis is to see if the rates of growth are different for novice as opposed to more experienced teachers. As done previously, the key results appear in Table 5.6 and the full results in the Appendix.

Table 5.6 Novice Model Results—Dimension-Specific Analysis

	CERR	ECL	MCP	MSB	CS	USDT	ESL	UAI
Fixed Effects (SE)								
Intercept (Score Beginning Y1)								
Experienced: β_{00}	**2.74**	**2.52**	**2.66**	**2.80**	**2.64**	**2.21**	**2.41**	**2.25**
	(0.03)	**(0.03)**	**(0.02)**	**(0.03)**	**(0.02)**	**(0.03)**	**(0.03)**	**(0.03)**
Δ Novice: β_{01}	**-0.32**	**-0.24**	**-0.23**	**-0.35**	**-0.24**	**-0.16**	**-0.25**	**-0.13**
	(0.06)	**(0.06)**	**(0.06)**	**(0.07)**	**(0.06)**	**(0.06)**	**(0.06)**	**(0.06)**
Slope (Growth per week)								
Experienced: β_{10}	-0.0009	-0.00043	0.0008	-0.0011	-0.001	-0.0004	0.0009	0.0008
	(0.0008)	(0.001)	(0.0008)	(0.0008)	(0.0009)	(0.001)	(0.0009)	(0.001)
Δ Novice: β_{11}	0.003	0.002	0.0003	0.004	0.002	0.0004	0.0027	-0.00045
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.0021)	(0.0019)
Variance Components								
Level 1: e_{ti}	0.27	0.31	0.29	0.23	0.29	0.34	0.3090	0.349
Level 2								
Intercept: r_{0i}	**0.08**	**0.08**	**0.07**	**0.12**	**0.07**	**0.07**	**0.07**	**0.08**
Slope: r_{1i}	**0.00003**	**0.00005**	0.00000	**0.00006**	**0.00004**	**0.00005**	**0.00005**	0.00002
N	458	458	458	458	458	458	458	458

*Statistically significant estimates ($p<0.05$) indicated in bold

The first key takeaway from Table 5.6 is all of the intercepts are statistically significant for novice and experienced teachers. Further, each novice coefficient is negative. This means that novice teachers receive lower scores on each dimension of the FFT, on average. Additionally, none of the growth parameters is statistically significant for novice or experienced teachers. Despite this, nearly every variance component is statistically significant. The only two that are not are the variance components that were non-significant in the unconditional growth model as well. These findings suggest that though there is variability both in the mean and growth scores across teachers for most dimensions, there is no empirical evidence to suggest that novice teachers have average rates of growth that differ from the growth rates of veteran teachers on any dimensions of the FFT.

The results presented thus far answer the first two research questions in the dissertation. We see that mean FFT scores range from about two to three on the FFT and there is variability in

these scores both within and across teachers. Growth trajectory estimates indicate that although there does not appear to be statistically significant growth over time, on average, the variance components indicate that some teachers do change, and that a few teachers increase or decrease by about half of an SD in FFT scores over the two years of the MET project. However, there is no difference in these growth rates between novice and experienced teachers. Further, there is no evidence to suggest that the rates of growth differ significantly across dimensions of the FFT for teachers in the MET project[19].

5.3 Gain Score Results

The analysis up until this point assumes teachers make salient changes in their practices both within and across years. The next set of results reflects an approach in which it is assumed that teachers only make measurable changes in their practices across years. For each teacher, all FFT scores within year are averaged together to give a mean score for each year. Thus, the growth parameter estimates in these models are gain scores between the first to the second year of the MET project.

[19] An interested reader can refer to Appendix E for a discussion regarding the assumptions of HLM and related limitations for this study, but a quick summary is that there are not concerns regarding any notable limitations to the growth estimates in the study due to the violation of assumptions of HLM.

Table 5.7 Gain Score Novice Model Results

Fixed Effects	**Coefficient**	**SE**
Intercept		
Experienced: β_{00}	2.51	0.02
Δ Novice: β_{01}	-0.22	0.04
Gain		
Experienced: β_{10}	0.01	0.02
Δ Novice: β_{11}	0.04	0.04
Random Effects	**Variance Component**	**p-value**
Level 1: e_{ti}	0.04	--
Level 2		
Intercept: r_{0i}	0.07	<0.001
Gain: r_{1i}	0.01	0.08
N	441	

Similar to the results in Table 5.1, the gain score results in Table 5.7 indicate that the intercept for both novice and experienced teachers as well as the variance component for the intercept are statistically significant. This means that there are statistically significant differences in the first year scores for novice as opposed to experienced teachers regardless of if growth in teacher practices is considered both within and across years or not. Additionally, there is significant variation among teachers in those differences. However, unlike in the growth parameter estimates, the gain score analysis does not yield statistically significant variance components for growth. In other words, gain scores between the two years of the MET project are not statistically significantly different from zero, on average. The estimated magnitude of the gain score is 0.01. The estimated weekly difference with the growth parameter estimates was -0.0002. Aggregating this value for an entire year indicates growth of about 0.01 as well, but in the negative direction. Both of these estimates were non-significant, on average, but it is helpful to know that their magnitudes were similar. This suggests that both the gain score and the longitudinal growth trajectories yielded similar measures of growth over time.

However, while the variance component was statistically significant with the growth trajectory estimates, there is no significant variability in the gain score estimates. As noted earlier, we expect those teachers in the tails of the growth parameter distribution to show growth up to 0.6 points in mean FFT score for an entire year. This means that assuming there is no growth within year can mask a time trend in teacher practices when those differences among teachers are small and that differences in growth might appear less than they actually are—particularly for the most extreme cases.

As has been shown thus far, there is little variability in growth across teachers in the MET project. Averaging scores across occasions to create gain scores, averaged away the small differences among teachers that were evident with the growth trajectories. These findings are consistent with prior literature regarding gain scores and the use of multiple time points to understand more about trajectories of growth rather than just differences between two time points (Rogosa et al., 1984).

Initial results indicate that growth parameter estimates that assume teachers change in their practices both within and across years might be superior to those that only assume change across years because statistically significant variability in growth might be lost when scores are averaged within year. However, it might be the case that the gain score estimates are more reliable measures of growth than the growth parameter estimates from the previous models (Rogosa & Willett, 1983). This point is investigated further in Section 5.5.

5.4 Growth Estimate Validity Results.

Before turning to an investigation of the reliability of the estimated growth parameters and gain scores, I include a brief validity check into the HLM growth parameter estimates as well as gain scores from the MET project. A comparison of the MET estimates as well as those from my independent analysis appear in Table 5.8. This table includes information regarding the number of videos each of the teachers submitted[20]; the MET and independent estimated growth parameters, year-specific mean scores, and gain scores; an indicator as to whether the direction of the MET growth parameters and independent estimates agreed; and a "Holistic Score". The holistic score indicates my thoughts on the direction of growth seen across the videos for that particular teacher over the course of the MET project. In other words, this is my one-word summary of if I believe I observed salient changes in teacher practices over the course of the submitted videos or not.

Table 5.8 Growth Estimate Comparisons

ID	# of Videos	Estimated Growth Parameter		Mean Y1 FFT score		Mean Y2 FFT score		Gain Score		Directional Agreement	Holistic Score
		MET	IND	MET	IND	MET	IND	MET	IND	Y/N	
1	8	0.0075	-0.0006	2.47	2.63	2.73	2.66	0.26	0.03	Y	Neither
2	15	0.0071	0.0135	2.70	2.45	3.12	3.06	0.49	0.61	Y	Improved
3	8	0.0071	0.0181	2.69	2.25	3.13	3.04	0.44	0.79	Y	Improved
4	8	0.0069	0.0123	1.91	1.88	2.50	2.38	0.59	0.50	Y	Improved
5	6	-0.0079	-0.0050	2.19	1.76	1.59	1.28	-0.59	-0.48	Y	Worsened
6	7	-0.0082	-0.0108	2.42	1.63	1.69	1.19	-0.73	-0.44	Y	Worsened
7	8	-0.0086	-0.0139	3.03	3.22	2.16	2.76	-0.88	-0.46	Y	Neither
8	8	-0.0094	-0.0086	2.22	2.38	1.56	2.04	-0.66	-0.34	Y	Worsened

The results in Table 5.8 indicate that the independent analysis of the videos agreed with the direction of growth for all eight of the extreme growth teachers, and that my overall assessment of teacher growth characterized by the holistic score was in agreement for all but two of eight cases. The HLM analysis indicated Teacher 1 was a positive high growth teacher.

[20] The MET data usage agreement does not allow for any demographic information be released when cell sizes are less than 10. Thus, no additional information about these teachers is provided.

Although I observed variability in this teacher's practices, it did not seem to me that those differences were signal of significant change in the teacher's behavior. Instead, I would argue that the differences among the lessons were more due to choice of lesson for each video rather than salient changes in the teacher's practices. Similarly, the HLM analysis indicated Teacher 7 was a negative high growth teacher. My assessment is that the first two lessons submitted by this teacher were of exceptional quality, but several of the factors that led to such high FFT scores were artifacts of the particular lessons. This teacher showed strong skills throughout the course of the MET project and I did not see indication that she actually worsened overall in her practices.

Differences in magnitude are likely due to several factors related to lack of validity of my own scores. That is, I did not complete any form of training with the FFT, nor did I pass a certification test, or have a MET researcher validate my application of the FFT on a regular basis. Additionally, due to time constraints, I scored all of these videos in blocks of time ranging from five to eight hours, as opposed to the four-hour rating shifts for MET raters. Rater fatigue definitely appeared in these long sessions and affected my ability to apply the FFT with consistent quality. Another difference is that I scored every video for every teacher in succession, while MET raters never scored more than one video per teacher per year. My scores certainly suffered from halo effects (when evidence from previous ratings affect subsequent ratings; Ho & Kane, 2013) in ways that MET rater scores could not have been affected. Finally, the issues with the server hosting the videos caused me to miss parts of most videos, which likely influenced scoring as well. In light of all of these issues with my own scoring practices, it is encouraging that my scores matched the direction of the HLM estimates for all of the teachers I observed, but my holistic disagreement with salient change for two teachers is cause for further study outside

of the scope of this project. I next turn to an investigation of the reliability of the growth estimates in each of the approaches and explore the conditions under which reliability is maximized in each.

5.5 Reliability of Growth Parameters

Although the application of HLM to observation scores is a novel aspect of this dissertation, a secondary, and equally valuable contribution of the project, is an investigation of the reliability of the estimated growth parameters. Since the MET project is time-unstructured and unbalanced, each teacher represents a separate measurement design with a unique estimate of growth and related reliability. Thus, the MET project provides a rich dataset for understanding how the measurement design influences the reliability of the growth parameters. An understanding of the reliability of growth parameters is useful because it informs future application of the methods demonstrated here. Recall from the Methods chapter that the reliability of a growth parameter is a function of three variables: variance of the growth parameter, variance of the error term, and the number and spacing of occasions. Understanding reliability necessitates understanding each of these components.

The least interesting of the three components that affect reliability is the variance of the growth parameters themselves. This is of least interest because it is the only factor that cannot be administratively controlled to some extent. Teachers are either variable in their growth rates or they are not. If it is the case that all teachers change (or do not) at close to the exact same rate, it is not easy to make distinctions among individuals, which is the purpose of increasing reliability. This is the case in the current project. As seen in the results presented in this chapter, the

variance of the growth parameters is very low (0.00004). Further, the ratio of this signal to noise is low. Thus, the reliability of the growth parameters are necessarily low in the current context.

Although the lack of variability (and poor signal to noise ratio) in observation scores is a limitation in reliably estimating growth trajectories in the current project, it is not a limitation to the application of the methodology described herein across all contexts. A district with more variability in observation scores due to the practices of their individual teachers (e.g. a context with higher signal to noise ratio) might have significantly more success in reliably modeling longitudinal growth trajectories in observation scores over time. The MET project data does not allow for investigating how more variation in observation scores affects reliability, but it does allow for an investigation of how measurement error and the number and spacing of occasions influences reliability. As such, the remainder of this section explores the relationship between each of these factors and the estimated reliability of the growth parameter both within and outside of the MET project context.

5.5.1 Reliability in the MET context. A first step in investigating reliability in the MET project is to gain an understanding of the various measurement designs present in the study. Figures 5.3 and 5.4 depict histograms for the distribution of the SST and related reliabilities respectively. The varying SSTs, or number and spacing of occasions, is what defines the different measurement designs. For reference, recall that in the unconditional growth model and novice model the variance of the growth parameter was 0.00004 and the variance of the level-1 error term was 0.14. These values combined with any SST from the distribution yields the distribution of reliabilities.

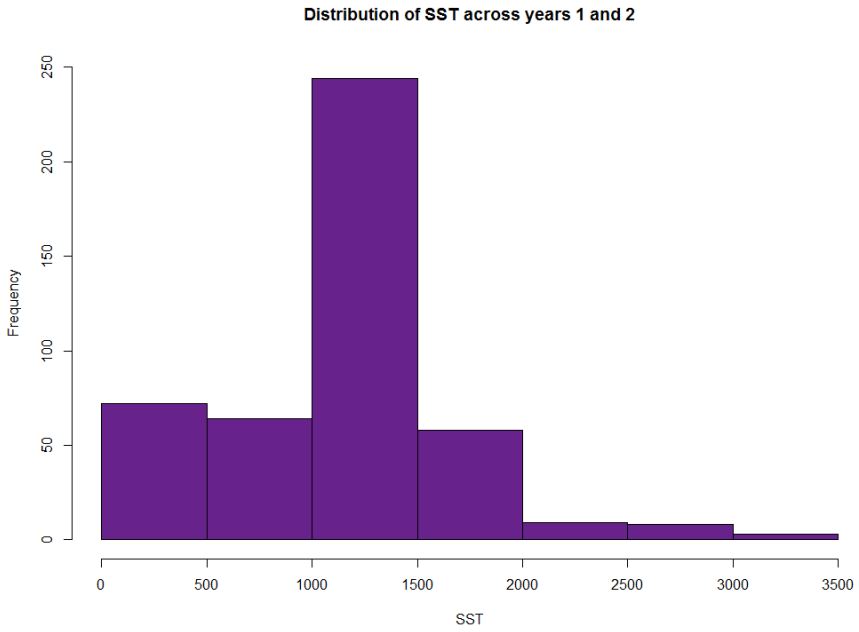


Figure 5.3 SST values in the MET project

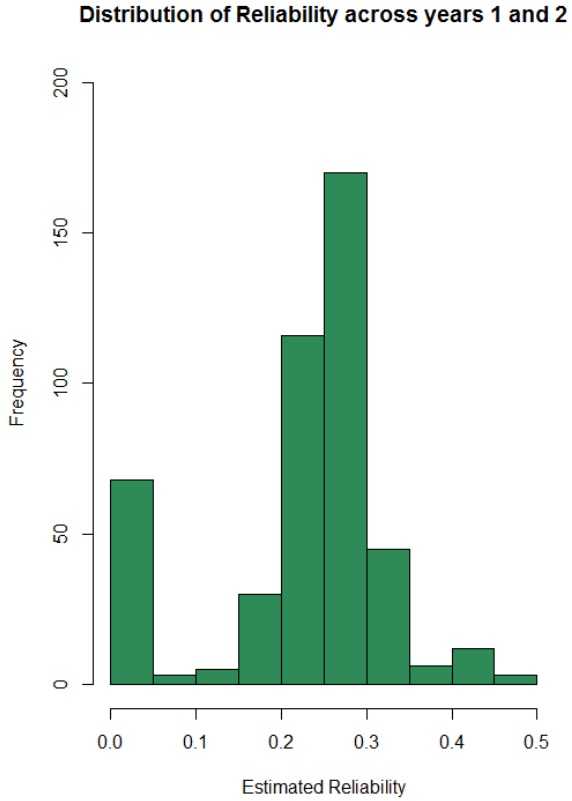


Figure 5.4 Reliability values in the MET project

The mean SST for the full dataset is 1113, and the standard deviation is 596. SSTs range from 0 to 3287. An SST of 0 results either from a measurement design with only one occasion or a design in which all of the occasions occur within the same week. The mean reliability for all of these designs is 0.22 with a standard deviation of 0.11. The average designs included about eight occasions with an SD of about 13 weeks between occasions. Reliability ranges from 0 to 0.48 with this dataset.

The design of the MET project allows for a helpful comparison between the SSTs and reliabilities for the first "year" of the project, which only spanned 19 weeks, and the second "year", which spanned 31 weeks. Teachers were to submit four videos per subject within each of these years. This means that the spacing of occasions was wider in the second year as compared to the first, but the number of occasions remained relatively stable. The longer spacing between the observations results in a higher value for SST, and this in turn mitigates the effect of error on reliability. A comparison between the two years of the data individually and combined are in Table 5.9.

Table 5.9 Comparison of SSTs and Reliabilities within MET Project Subsets

	Year 1 Four occasions; Nineteen Weeks	Year 2 Four occasions; Thirty-one Weeks	Both Years Eight occasions; Fifty Weeks
SST			
Mean	22	266	1113
Median	17	255	1205
Standard Deviation	23	114	596
Minimum	0	0	0
Maximum	146	860	3287
Reliabilities			
Mean	0.01	0.02	0.22
Median	0.01	0.02	0.25
Standard Deviation	0.01	0.01	0.11
Minimum	0.00	0.00	0.00
Maximum	0.07	0.05	0.48
N	456	392	458

The results in Table 5.9 indicate what we expect given the nature of the reliability formula. Reliability increases when the spacing of the observations increases (shift from 0.01 to 0.02), but the increase is even larger when both the number and spacing of occasions increase (shift from 0.01 to 0.22). The relative magnitudes of these shifts suggest that adding more occasions has a higher influence on increasing reliability than increasing the spacing.

In order to gain a better understanding of the relative influence on both the number and spacing of occasions on SST. Table 5.10 provides information regarding the relationship between the number of occasions, the SD of the weeks in which occasions took place, and SST. These results come from regressing number of occasions and SD of occasions on SST individually in two separate simple linear regressions and then again together in a multiple linear regression for both years of data. The mean and SD of the number of occasions are 7.36 and 2.70 respectively, and the mean and SD of the standard deviation between occasions are 12.14 and 3.88 weeks respectively.

Table 5.10 Relative Impact of Number and Spacing of Occasions on SST

	Unstandardized Coefficient	Standardized Coefficient
Simple Linear Regression		
Number of Occasions	196.33	0.89
SD of Occasions	117.85	0.81
Multiple Regression		
Number of Occasions	140.5	0.61
SD of Occasions	66.2	0.45

The standardized coefficients from the simple linear regressions suggest that the number and spacing of occasions have relatively equal influence on SST. However, the results from the multiple regression in the bottom panel of Table 5.10 indicate that once both variables are included in the regression, the impact of the two variables is less similar. While both number and

spacing of occasions still have significant influence over SST, the number of occasions clearly has a stronger influence. In other words, increasing the number of occasions has a greater impact on increasing reliability than increasing the spacing between occasions. The idea that more occasions significantly increases reliability is intuitive, but it is notable that even after controlling for the effect of the number of occasions, the spacing of those occasions is still a significant predictor of SST. This, in turn, means that the spacing of occasions has important influence on reliability.

The number and spacing of occasions and their relationships to SST have clear implications for the reliabilities of growth parameters. The results thus far suggest that adding more occasions is more beneficial than increasing the spacing between occasions. A closer look at just those measurement designs in the top quartile of the SST distribution helps to elucidate even more information on this point.

The results in Table 5.9 indicate that the mean reliability for all designs in the MET project is 0.22. However, the mean reliability for the top quartile of SSTs is 0.32. Recall that the MET project design included four videos per year per teacher for each subject. For elementary school teachers, this might mean a total of 16 videos over the course of the study. It might be the case that all of the measurement designs in the top quartile of the distribution included 16 occasions over the two years. As the results in Table 5.10 suggest, more occasions drives higher reliabilities, so this makes sense. It is important to know if there are designs with fewer occasions in this top quartile as eight observations per teacher per year is costly and unrealistic for most schools and districts. Figure 5.5 provides the distribution of the number of occasions for the models present in the top quartile of SST values. The mean number of occasions within this subset is 9.5 and the mean SD of the spacing of occasions is 14.6 weeks (as compared to 7.36

and 2.7 respectively in the full sample). This means that these teachers tend to have 9-10 occasions spread over the course of about 14 months, or about five occasions per school year in the study. The range of reliabilities for these designs runs from 0.28 to 0.48.

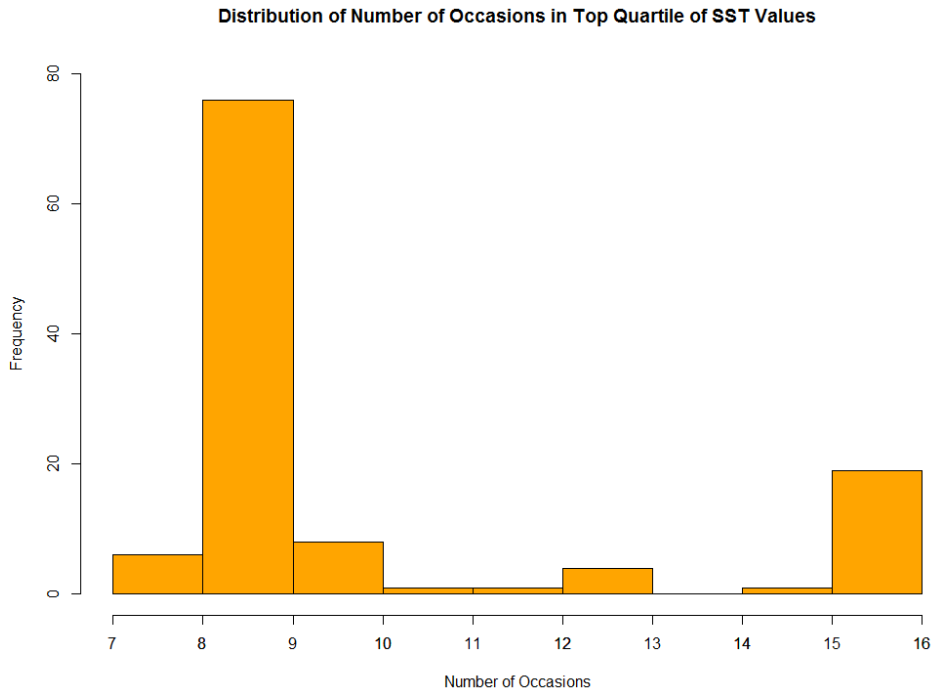


Figure 5.5 Number of occasions for measurement designs yielding highest reliabilities

The measurement designs in the top quartile of SST distributions have a mean reliability of 0.32, and the design with the highest reliability is about 0.5. In other words, about half of the variation in scores is still due to noise rather than signal in the most reliable model in the MET project. However, previous research with value-added models (another aspect of measuring teacher quality) both within and outside of the MET project report reliabilities ranging from 0.3 – 0.5 (Kane & Staiger, 2012; McCaffrey et al., 2009). The mean reliability of the growth parameter is low in the current study, but the reliabilities for the measurement designs in the top quartile of SSTs fall within the range of reliabilities for VA scores, albeit on the lower end of the range we see there. This helps to provide context for the current levels of reliability; although they are

modest, they are not outside the range of reliabilities for other growth-based measures widely used in teacher evaluation systems.

Although the reliabilities present in the top quartile of the distribution are in line with what we see in value-added models, on average, they are still not ideal. However, there is another way to increase the reliability of growth parameters aside from the number and spacing of occasions. In the next section, I consider how adding a second rater to each observation might change the reliabilities seen in the current context. In addition to the number and spacing of occasions, the number of raters is another way decision-makers can indirectly influence the reliability of growth parameter estimates.

5.5.2 Reliability of growth parameters outside the MET context. As discussed in the previous chapters, the reliability of FFT scores for any given occasion can be increased either by averaging over multiple occasions or adding additional raters to a given occasion. Since all occasions are necessary for estimating growth parameters, averaging over multiple lessons does not make sense in the current project. However, measurement error due to rater for individual scores could be reduced by adding another rater per occasion. For instance, if one rater becomes distracted by a passing thought while viewing a particular lesson, it is probable that another rater will not experience that same distraction. Thus, the presence of scores from both raters would provide a more accurate indication of the teacher's practices at any given occasion. If the scores are more reliable for each occasion, it is not unreasonable to then expect the reliability of the growth rate to improve as well. Unfortunately, there is no way to know exactly how much adding a second rater to every lesson in the MET project would affect measurement error and reliability of growth parameter estimates, but prior research provides some insight into a reasonable thought experiment about what might happen.

As a part of an extension of the MET study, Ho & Kane (2013) investigated how the reliability of FFT scores on a given occasion changed depending on who provided the rating scores (e.g. administrator vs. peer rater) or how many raters observed each teacher. The study included 67 teachers with four lessons each for a total of 268 videos. The rater pool included 53 school administrators and 76 peer raters, for a total of 129 raters. In order to mitigate the rating load, the study did not use a fully crossed design in which each rater scored every video. Instead, raters scored four lessons from each of six different teachers for a total of 24 lessons per rater. Videos were fully crossed within each team of raters, and then Ho & Kane pooled scores across blocks in order to identify variance components in the observation scores. An important component of this study is that Ho & Kane considered lessons nested within teachers. Because of this, adding an additional rater to a teacher is synonymous with adding an additional rater for a lesson. Ho & Kane's findings indicate that shifting from one to two raters per lesson can increase the reliability of observation scores from that occasion by about 23% (p. 15).

Ho & Kane's research suggests that moving from one to two raters increases the individual reliability of FFT scores by 23%. This also means there is a 23% decrease in the mean-squared error for any given occasion. I extend this decrease in measurement error to the context of the reliability of the growth parameters by assuming that decreasing the measurement error of all the occasions used to estimate a growth trajectory also reduces the variance of the level-1 error term of the growth trajectory by the same magnitude.

Table 5.11 includes four different sample measurement designs from the MET project: two with four observations over two years and two with eight observations over two years. Within each context (four or eight observations total), there is one case in which the SST is maximized and another in which SST is minimal. The first row of Table 5.11 provides the

reliabilities of each of these four measurement designs from the MET project. The second line indicates the reliability of the growth parameter if measurement error is reduced in each context by 23%, and the bottom row of Table 5.11 indicates the proportion of increase in reliability in going from one to two raters for each measurement design.

Table 5.11 Reliability of Growth Parameters in Best and Worst Case Designs

	4 observations over 2 years		8 observations over 2 years	
	Minimal Spacing $SST_{41} = 34$	Maximal Spacing $SST_{42} = 284$	Minimal Spacing $SST_{81} = 1246$	Maximal Spacing $SST_{82} = 2171$
1 Rater	0.0096	0.075	0.263	0.383
2 Raters	0.0124	0.095	0.316	0.446
% Increase 1-2 Raters	29	27	20	16

Adding a second rater when there are only four occasions increases reliability by about 30%, but when there are eight occasions, the additional rater only increases the reliability by about 20%. The results in Table 5.11 indicate that adding a second rater has greater impact on reliability when the spacing is mediocre or there are fewer occasions. In other words, adding occasions is the most effective way of increasing the reliability of the growth parameter estimate, followed second by maximizing the spacing between occasions and then finally adding a second rater.

This investigation of reliability provides more information on the distribution of reliabilities in the MET project and details regarding some best and worst case scenarios for estimating growth in teacher practices over time. The designs with eight or more occasions and greater spacing have higher levels of reliability. Adding additional raters to a measurement design increases reliability, but the impact of a second rater is not as high when there are more occasions. Unfortunately, eight occasions a year with a single rater is a significant burden for

many districts, let alone eight occasions with multiple raters. Regardless, the discussion of reliability included here allows for understanding of how the choices around number and spacing of occasions as well as number of raters affects reliability. A key implication is that it seems likely that designs with only two occasions per year would not be able to provide reliable estimates of growth, assuming the true growth variance is very small as appears to be the case in the MET project. This is key information for designing observation systems if the use of growth estimates is a potential outcome of interest.

5.5.3 Reliability of gain scores. Unlike the growth parameter estimates discussed so far, the gain score analysis only has one reliability. This is because there are no longer multiple measurement designs in play. Instead, there is one way to estimate a gain score in the current study. The reliability of β_{10} from HLM7 is about 0.10. Recall that the average reliability for the growth parameter estimates was 0.22. Thus, at first consideration, it appears that the growth parameters assuming growth both within and across years yield more reliable estimates of growth than gain scores. However, similar to the previous investigation of best case scenarios for growth within and across years, a similar thought experiment can be considered for the reliability of gain scores.

In the current analysis, the gain scores are based on FFT scores with only one rater per occasion, and four occasions per year. We could further minimize measurement error both by adding more occasions and by including additional raters per occasion. The previously discussed article by Ho & Kane (2012) offers evidence to suggest that an observation system including six different occasions with six different raters yields reliability estimates for status FFT scores of

0.72. I use this as a model for a best-case scenario and to investigate the reliability of a gain score with such a measurement design.

Using the variances, standard deviations, and correlations from the current gain score analysis along with the implied reliability from the Ho & Kane study, we can get an estimate for the reliability of gain scores based on six observations per year, each with a different rater.

$$\rho(D) = \frac{0.110*0.72+0.118*0.72-2*0.331*0.344*0.572}{0.110+0.118-2*0.331*0.344*0.572} = \frac{0.0339}{0.0977} = 0.35$$

These results indicate that a design as optimistic as Ho & Kane's still only yields a reliability of 0.35 for gain scores across years. If we go even further and imagine a scenario with 2-3 raters on each of 5 different occasions in a year, where the reliability of mean scores within year may reach as high as 0.8, the resulting gain score reliability is still only 0.53.

Recall that the reliability analysis for the growth parameter estimates indicated that a measurement design including four occasions per year, with only one rater per occasion, but maximal spacing between occasions yielded a reliability of 0.38, and adding a second rater implied a reliability of about 0.45. Further, there was one measurement design in the MET project with 8 occasions in each year that yielded a reliability of 0.48. What this final analysis shows is that the reliability of growth parameter estimates is better with more, less-reliable observations as a basis for a growth estimate than having fewer, more-reliable estimates. Although it is possible to get higher reliability with a gain score analysis, the resources necessary for such outcomes are much more costly than those with growth parameter estimates.

5.6 Conclusion

Although there is evidence in this study that FFT scores change over time, there is no empirical evidence to suggest that the rate of change differs based on a teacher's experience level. Thus, the current project provides little evidence to support the hypothesis that novice teachers change more in their classroom practices than veteran teachers. It is possible that the difference does not exist. However, it is also possible that the difference between the groups is small and the measurement designs in the MET project were not powerful enough to detect such a small difference in rates of change.

Increasing the reliability of growth estimates makes it more possible to detect small differences in rates of change. The close analysis of the relationship between SST and reliability of growth parameter estimates presented here indicate that additional occasions have the greatest impact on reliability, followed next by increased spacing between occasions, and finally by adding an additional rater. Further, it is easier to gain higher levels of reliability if growth is assumed to occur both within and across years. Although it is possible to attain higher levels of reliability with gain scores, the cost for the necessary observation system is great. Local education agencies should take care to consider each element of an observation system and how it potentially affects the reliability of the growth parameter if it will be used for decision making of any kind. Further discussion of these results in the current teacher evaluation policy context is in the next, and final, chapter of this dissertation.

6.0 Discussion

Prior research suggests that teachers are dynamic in their practices and that change occurs at all stages in teachers' careers. There is also evidence to suggest that teachers in their first few years of teaching change the most in their practices as they learn from their early classroom experiences. Particularly, novice teachers are more variable in their practices related to responsibilities such as managing student behavior or managing classroom practices. However, in this study, the variance components from the HLM analysis suggest that there is growth in FFT scores over time for teachers, but there are no findings to indicate that the rate of growth is different for novice teachers as compared to more experienced teachers. In addition, there is no evidence to suggest that novice teachers grow more quickly in some practices than others.

The final analysis in the current project focused on the reliability of the growth parameter estimates. Although the mean reliability in the current context was low, reasonable levels can be reached with enough occasions and raters. Comparing the results across measurement designs and how reliability might be increased are the most important findings from this section in an applied policy context. This analysis provides information necessary for designing an observation system. Those in charge of designing observation systems should know that the best way of increasing the chances of identifying differences in rates of change among teachers is to observe teachers frequently, then to make sure that those occasions are widely spaced, and, if possible, to have multiple individuals observe on any given occasion. Decision-makers should also be aware of the reliability of individual observation scores. Although the reliability of status FFT scores can be increased by averaging across occasions, doing this limits the reliability of growth estimates.

As a final general comment, it is important to remember the context of the MET project. Specifically, there were no stakes associated with these observations and there was no personal relationship between teachers and raters. In some ways, this might be considered a best case scenario for observation scores. That is, the raters had no personal investment in the outcome of the observation scores, and the raters felt no pressure that their scores would be used to make high-stakes decisions. If there were stakes attached to growth in observation scores, raters might use the highest scoring categories more frequently and less in the lower categories. Then even less growth might be evident. If showing growth in teacher practices over time is something that is highly valued, this is an important concern that should be addressed when designing the observation system.

6.1 Limitations and Future Research

The results from this dissertation illustrate that it can be difficult to reliably model growth in observation scores. However, some of that might be due to the context of this particular dataset. The generalizability of the current findings are relatively limited. Not only are the districts present in the MET project not a random sample from the United States, but the sample of teachers used in this study are not even a random sample from within the MET project data. About one-third of teachers in the MET project did not have information regarding years of experience and another 40% of teachers did not have dates associated with their videos. None of these teachers could be used in the analysis. As such, the findings here cannot be generalized to the full population of the MET project let alone to any other greater population of teachers.

A dataset that chronicled teachers' observation scores for longer periods of time, perhaps three - five years, and with a more robust sample of teachers might be able to reveal more information about the ways teachers change over time. Although many districts may not have up to eight occasions per elementary teacher per year, there are districts that have consistently collected at least two occasions per year for several years for at least some teachers. The additional years of data will surely increase the reliability of growth parameter estimates, even with fewer occasions within year. Thus, the methods described in this dissertation should be implemented in a district context where all occasions can be accurately mapped to specific dates and complete information is available regarding the number of years of teacher experience for multiple years. Such a study would provide more information about what might be expected in other contexts and if the results in the current project are anomalous due to the context of the MET project.

Additionally, the requirements for higher reliabilities presented in this study may be untenable in many local contexts. Recorded lessons certainly make the logistics of scoring multiple lessons per teacher easier than live observations, but scoring still requires significant person-hours. Further, as pointed out in this project, video scoring can limit the ability to apply full observation protocols. Additionally, recorded videos require districts to have sufficient recording devices to collect video data at a reasonable pace for all teachers. More work should be done combining the methods here with prior work regarding the length of observations (c.f. Ho & Kane, 2012; Mashburn et al., 2014). Multiple observations of only 20 minutes per observation requires fewer resources than observations lasting full class periods, but work is needed to understand the influence of observation length on the reliability of growth parameter estimates.

Finally, given the low reliability of the growth estimates, it is difficult to recommend growth measures be used for making decisions about individual teachers, but this does not limit the potential use of growth scores for research purposes. Although this study focuses on the reliability of growth parameter estimates, it is possible that growth scores might be used in ways that rely more heavily on precision rather than reliability. Here, I refer to the standard error of measurement as a quantification of precision. The standard error of measurement provides information regarding how accurately a single score approximates the expected value for a particular individual (Haertel, 2006). Reliability is of most importance when rank ordering individual teachers is of great importance (Rogosa et al., 1982), but if we are more interested in meeting a particular thresh hold, precision may be of more value. For example, a district may expect to see a specific amount of growth in teacher practices after a new teacher induction or professional development program implementation. Rather than valuing the order of magnitude for each teacher's growth, the metric of interest may be the proportion of teachers who grow at least by a predetermined amount. In this instance, the precision of the growth estimates would be more valuable than the reliability. Further, rather than using the growth estimates to evaluate particular teachers, the proportion of teachers who reached the desired growth rate might better inform the district of the value of the induction or professional development program.

6.2 Conclusion

Teacher observations have changed in multiple ways in the past decade. Not only has more attention been paid to the quality of the observation protocols, but more consideration has also been put into how these protocols are implemented. For example, more and more districts

require certification training for all administrators who will score teachers with a particular rubric. Additionally, the number and frequency of observations is expanding in multiple contexts. A byproduct of these new observation systems is the collection of richer longitudinal data designed to measure the quality of teachers' practices.

Using observation scores to understand the longitudinal trajectories of teacher practices is at its infancy. No prior work has investigated the use of observation scores over time. As such, it is important that care be used in understanding this use of observation scores before it is taken up in any widespread way. This project implements a first approach at understanding changes in teacher practices over time, and provides helpful information for individuals in charge of designing observation systems. Although it is not advisable to take this dissertation as evidence for including growth in teacher evaluations directly, estimating longitudinal growth trajectories for teachers can have multiple positive uses. Identifying induction and professional development programs that help teachers improve classroom practices in objective ways would not only build the research literature on learning to teach but also help local education agencies select proven programs for their educators. In addition, data showing the general trends for a particular district context would help to inform more useful professional development activities for particular populations of teachers. These broad uses of growth parameter estimates are helpful for both research and practitioner communities without placing high stakes on these relatively unreliable measures.

Schools and districts dedicate a good deal of effort and resources to helping teachers improve in their practices. Coaching, professional development seminars, lesson studies, and peer mentoring require money, time, and personnel. The motivation behind these expenditures is the expectation that they make a difference in the classroom practices of teachers. Ideally, these

would be long-lasting and positive changes. Aside from these intentional efforts, it is logical to expect that professionals, regardless of discipline, to experience varying levels of success over the course of their careers. Being able to identify the quality of a teacher's practices as well as change in that quality over time is helpful information previously unavailable for research or program evaluation purposes.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Atteberry, A., Loeb, S., & Wyckoff, J. (2017). Reassignment rates and implications for student achievement. *Educational Evaluation and Policy Analysis*, *39*(1), 3-30.

Beck, C., Kosnik, C., & Rowsell, J. (2007). Preparation for the first year of teaching: Beginning teachers' views about their needs. *The New Educator*, *3*(1), 51-73.

Berliner, D. C. (2004). Describing the behavior and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society*, *24*(3), 200-212.

Bill and Melinda Gates Foundation. (2013). Measures of Effective Teaching: 1 - Study Information. ICPSR34771-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2013-09-23. http://doi.org/10.3886/ICPSR34771.v2

Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. E. (2008). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education*, *24*(2), 417-436.

Braun, H. I. (2005). Using student progress to evaluate teachers : A primer on value-added models. Princeton.

Briggs, D. C. (2012). Making value-added inferences from large-scale assessments. *Improving large-scale assessment in education: Theory, Issues and Practice*, 186-206.

Cantrell, S., & Kane, T. J. (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study. *Policy and Practice Brief*. MET Project. *Bill & Melinda Gates Foundation*.

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

Cronbach, L. J., & Furby, L. (1970). How we should measure" change": Or should we? *Psychological Bulletin*, *74*(1), 68.

Danielson, C. (2011). *The framework for teaching: Evaluation instrument*. Princeton, NJ: Danielson Group.

Danielson Group. (2013). Charlotte Danielson. Retrieved from https://www.danielsongroup.org/charlotte-danielson/

Doherty, K. M., & Jacobs, S. (2015). State of the states 2015: Evaluating teaching, leading and learning. National Center on Teacher Quality. *National Council on Teacher Quality.* http://www. nctq. org/dmsView/StateofStates2015.

Doherty, K. M., & Jacobs, S. (2013). State of the states 2013: Connect the dots: Using evaluations of teacher effectiveness to inform policy and practice. *National Council on Teacher Quality*. http://www. nctq. org/dmsView/State_of_the_States_2013_Using_Teacher_Evaluations_NCTQ_Report

Feiman-Nemser, S. (1983). Learning to teach. *ERIC Digest.* Retrieved from http://www.eric.ed.gov

Feiman-Nemser, S. (2001). From preparation to practice: Designing a continuum to strengthen and sustain teaching. *Teachers College Record*,*103*(6), 1013-1055.

Forte, E., edCount, L. L. C., Perie, M., & Paek, P. (2014). Exploring the relationships between English Language Proficiency Assessments and English Language Arts Assessments. *Center for Assessment, US Department of Education*.

Franke, M. L., Carpenter, T. P., Levi, L., & Fennema, E. (2001). Capturing teachers' generative change: A follow-up study of professional development in mathematics. *American Educational Research Journal*, *38*(3), 653-689.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915-945.

Gordon, R. J., Kane, T. J., & Staiger, D. (2006). *Identifying effective teachers using performance on the job*. Washington, DC: Brookings Institution.

Grossman, P. L., Valencia, S. W., Evans, K., Thompson, C., Martin, S., & Place, N. (2000). Transitions into teaching: Learning to teach writing in teacher education and beyond. *Journal of Literacy Research*, *32*(4), 631-662.

Haertel, E. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (65 – 110). Westport, CT: American Council on Education and Praeger Publishers.

Harootunian, B., & Yarger, G. P. (1981). Teachers' conceptions of their own success. Current Issues.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education*, *4*(4), 319-350.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64. http://doi.org/10.3102/0013189X12437203

Ho, A. D., & Kane, T. J. (2013). The reliability of classroom observations by school personnel. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

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*.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*(4), 54-78.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). Guilford publications.

Lavigne, A., & Bozack, A. (2015). Successes and struggles of teaching: Perspectives of beginning, mid-career, and veteran teachers. *Journal of Teaching Effectiveness and Student Achievement*, *2*(2), 68-80.

Lumpe, A., Czerniak, C., Haney, J., & Beltyukova, S. (2012). Beliefs about teaching science: The relationship between elementary teachers' participation in professional development and student achievement. *International Journal of Science Education*, *34*(2), 153-166.

Martin, S. D. (2004). Finding balance: Impact of classroom management conceptions on developing teacher practice. *Teaching and Teacher Education*, *20*(5), 405-422.

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement*, *74*(3), 400-422.

McDonald, F. (1980). The problems of beginning teachers: A crisis in training. *Study of induction programs for beginning teachers*. Princeton, New Jersey: Educational Testing Service, 1980.

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

McClellan, C., Donoghue, J., & Park, Y. S. (2013, April). Commonality and uniqueness in teaching practice observation. In *annual meeting of the National Council of Measurement in Education, San Francisco, CA*.

Milanowski, A. (2011). Strategic measures of teacher performance. *Kappan Magazine*, *92*(7), 19-26.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33-53.

National Board Resource Center Illinois State University. (2017). Charlotte Danielson Framework for Teaching aligned to the National Board's Five Core Propositions. Retrieved from http://nbrc.illinoisstate.edu/downloads/nbrc/crosswalk/1charlottedanielson.pdf

New York City Department of Education. (2017). Alignment across the NYCDOE: Linking each element of the Framework for Great Schools with NYCDOE measures and resources. Retrieved from http://schools.nyc.gov/NR/rdonlyres/7D5834A8-A01D-4D99-9F28-6E0D613EBC69/0/FrameworkforGreatSchoolsAlignmentAcrosstheNYCDOE.pdf

Peressini, D., Borko, H., Romagnano, L., Knuth, E., & Willis, C. (2004). A conceptual framework for learning to teach secondary mathematics: A situative perspective. *Educational Studies in Mathematics*, *56*(1), 67-96.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning?. *Educational Researcher*, *29*(1), 4-15.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (Vol. 1). Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & Tolt, M. (2011). *HLM 6: Hierarchical linear and nonlinear modeling*. Scientific Software International.

Rogosa, D., Floden, R., & Willett, J. B. (1984). Assessing the stability of teacher behavior. *Journal of Educational Psychology*, *76*(6), 1000-1027.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability the difference score in the measurement of change. *Journal of Educational Measurement*, *20*(4), 335-343.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*(3), 726.

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175–214. http://doi.org/10.1162/qjec.2010.125.1.175

Ryan, K. (Ed.). (1970). Don't smile until Christmas: Accounts of the first year of teaching. University of Chicago Press.

Singer, J. D., & Willett, J. B. (2003). Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford university press.

Snyder, C. (2012). Finding the" royal road" to learning to teach: Listening to novice teacher voices in order to improve the effectiveness of teacher education. *Teacher Education Quarterly*, *39*(4), 33-53.

Office of Superintendent of Public Instruction. (2017). State of Washington. Danielson Framework for Teaching aligned with the Washington Eight Teacher Evaluation Criteria. Retrieved from http://www.k12.wa.us/TPEP/Frameworks/Danielson/Danielson_WA_Alignment.pdf

Thompson, J., Windschitl, M., & Braaten, M. (2013). Developing a theory of ambitious early-career teacher practice. *American Educational Research Journal*, *50*(3), 574-615.

U.S. Dept. of Education. (2009). *Race to the Top Program: Executive Summary*. U.S. Department of Education Retrieved from https://www2.ed.gov/programs/racetothetop/executive-summary.pdf.

Wang, J., Odell, S. J., & Schwille, S. A. (2008). Effects of teacher induction on beginning teachers' teaching: A critical review of the literature. *Journal of Teacher Education*.

Watzke, J. L. (2007). Foreign language pedagogical knowledge: Toward a developmental theory of beginning teacher practices. *The Modern Language Journal*, *91*(1), 63-82.

Webb, N. L. (1997). Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education. Research Monograph No. 6.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. New York, NY.

Whitworth, B. A., & Chiu, J. L. (2015). Professional development and teacher change: The missing leadership link. *Journal of Science Teacher Education*, *26*(2), 121-137.

Willett, J. B. (1997). Measuring change: What individual growth modeling buys you. In E. Amsel and K. Renninger (Eds.), *Change and development: Issues of theory, method, and application*, (213- 243). Winnipeg: International Institute for Sustainable Development.

Willett, J. B. (1994). Measurement of Change. In T. Husen and T. N. Postlethwaite (Eds.), *International Encyclopedia of Education,(second ed.) (*671-678). Oxford: Pergamon Press.

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement*, *49*(3), 587-602.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education, 15*, 345-422.

Appendices

A: Confirmatory Factor Analysis Estimation Details

The confirmatory factor analysis (CFA) conducted in order to better understand the latent structure of the FFT observation protocol occurred with the lavaan R statistical package. The method of estimation was diagonalized weighted least squares (estimator = "DWLS") with robust standard errors (se = "robust"). Additionally, mean and variance adjusted test statistics were used (test= "scaled.shifted") as well as fixed variances for each latent variable (std.lv=TRUE). The two-factor model estimated a separate factor for the two domains of the FFT (i.e. instruction and environment), with the corresponding four items loading on each factor as specified by Danielson (2011). The one-factor model constrained the correlation between the two factors to 1, so that the one-factor model was nested within the two-factor.

B: Danielson Framework Abbreviated Scoring Rubrics

Dimension 1: Creating an environment of respect and rapport	
1	Unsatisfactory—negative and inappropriate interactions, teacher does not respond
2	Basic—appropriate interactions but inconsistent, teacher tries to respond
3	Proficient—friendly and respectful interactions, teacher responds successfully
4	Distinguished—highly respectful interactions, students and teacher respond successfully
Dimension 2: Establishing a culture for learning	
1	Unsatisfactory—lack of commitment to learning, no investment of student energy into tasks
2	Basic—little commitment to learning, students complete tasks without high quality
3	Proficient—commitment and value to learning, students engage in tasks
4	Distinguished—shared commitment to learning, students engage in tasks with high quality
Dimension 3: Managing classroom procedures	
1	Unsatisfactory—instructional time is lost, inefficient classroom routines
2	Basic—some instructional time is lost, partially effective classroom routines
3	Proficient—little loss of instructional time, consistently follow established classroom routines
4	Distinguished—instructional time is maximized, students contribute to management and classroom routines
Dimension 4: Managing student behavior	
1	Unsatisfactory—little monitoring of student behavior, repressive or disrespectful response
2	Basic—tries to monitor student behavior, inconsistent response
3	Proficient—monitors student behavior, consistent and successful response
4	Distinguished—teacher and students monitor own behavior, sensitive response to individual needs

Figure B.1 FFT Domain 2 scoring rubrics. Adapted from DS4: Framework for Teaching (FFT) – Year 2, by Bill and Melinda Gates Foundation, 2014.

Dimension 1: Communicating with students	
1	Unsatisfactory—unclear purpose, confusing directions, content errors, incorrect vocabulary
2	Basic—partial purpose, clarified directions, minor content errors, limited vocabulary
3	Proficient—clear purpose, clear directions, accurate content, appropriate vocabulary
4	Distinguished—link purpose to interests, clear directions, thorough content, extends vocabulary
Dimension 2: Using questioning and discussion techniques	
1	Unsatisfactory—low cognitive challenge, single correct responses, few students dominate
2	Basic—moderate cognitive challenge, thoughtful responses, engages all students
3	Proficient—cognitive challenge that advances thinking, genuine discussion with all students
4	Distinguished—promotes meta-cognition, high-level thinking, students formulate discussion
Dimension 3: Engaging students in learning	
1	Unsatisfactory—poorly aligned lessons, rote responses, no structure, slow/rushed pace
2	Basic—partially aligned, minimal thinking responses, recognizable structure, rushed pace
3	Proficient—aligned, challenges thinking, clear structure, appropriate pacing
4	Distinguished—fully aligned, students initiate inquiry, clear structure, good pacing and choices
Dimension 4: Using assessment in instruction	
1	Unsatisfactory—little monitoring of student learning, absent feedback, unaware of assessment criteria
2	Basic—some monitoring of student learning, general feedback, partially aware of assessment criteria
3	Proficient—regular monitoring of student learning, specific feedback, aware of assessment criteria
4	Distinguished—regular monitoring of student learning, fully integrated feedback and assessment criteria

Figure B.2 FFT Domain 3 scoring rubrics. Adapted from DS4: Framework for Teaching (FFT) – Year 2, by Bill and Melinda Gates Foundation, 2014.

C: Data Creation

Data cleaning and coding for this project occurred in R version 3.2.2 through the Inter-university Consortium for Political and Social Research Virtual Data Enclave. This is the secure access platform provided to anyone wishing to access to the MET data. The cleaned data used in this project came from multiple raw data files within the MET database. Table C.1 provides a list of the variables selected from each of the respective raw data files in order to create the data used for this study.

Each of the MET raw data files was subsetted to include only the variables in Table B.1. The files were then merged based on whatever combination of DISTRICT_ICPSR_ID, SCHOOL_ICSPR_ID, TEACHER_ICPSR_ID, SECTION_ICPSR_ID, VIDEO_ICPSR_ID, GRADE, and SUBJECT variables available in the individual merges. Finally, CAPTUREDATETIME was formatted as a character variable, so it was reformatted and saved as a new variable called "date". The final data included all of the variables in the first four columns of Table C.1 plus the additional variables listed in the last column.

Table C.1 Data Sources and Final Variables for Analysis

FFT File	Dating File	Teacher File	Section File	Final Data
DISTRICT_ICPSR_ID	DISTRICT_ICPSR_ID	DISTRICT_ICPSR_ID	DISTRICT_ICPSR_ID	FFT_Total
SCHOOL_ICSPR_ID	SCHOOL_ICSPR_ID	SCHOOL_ICSPR_ID	SCHOOL_ICSPR_ID	Date
TEACHER_ICPSR_ID	VIDEO_ICPSR_ID	TEACHER_ICPSR_ID	TEACHER_ICPSR_ID	Day
SECTION_ICPSR_ID	GRADE	DAD_MALE	SECTION_ICPSR_ID	Week
VIDEO_ICPSR_ID	SUBJECT	DAD_WHITE	YEAR	Occasion
RATERID	CAPTUREDATETIME	DAD_BLACK	GRADE	
YEAR		DAD_HISPANIC	DAD_PERCGIFTED	
GRADE		DAD_YRSEXP	DAD_PERCMALE	
SUBJECT		DAD_YRSEXPDIST	DAD_PERCSPED	
FFT_CERR		DAD_MASTERSPLUS	DAD_PERCELL	
FFT_USDT			DAD_PERCFRL	
FFT_ECL			DAD_PERCHISPANIC	
FFT_MCP			DAD_PERCBLACK	
FFT_CS			DAD_PERCWHITE	
FFT_MSB				
FFT_ESL				
FFT_UAI				

D: Full HLM Results

This appendix provides the full results from the dimension-specific HLM analysis. The key results most relevant for discussion are abbreviated and provided in the Results chapter. Statistically significant estimates ($p<0.05$) indicated in bold in all tables.

Table D.1 Unconditional Means Model Dimension-Specific Fixed Effects Estimates

Dimension	Coefficient	Standard Error	t-ratio	Degrees of freedom	p-value
CERR	**2.67**	**0.02**	**156.79**	**457**	**<0.001**
ECL	**2.46**	**0.02**	**143.64**	**457**	**<0.001**
MCP	**2.64**	**0.02**	**158.55**	**457**	**<0.001**
MSB	**2.74**	**0.02**	**162.72**	**457**	**<0.001**
CS	**2.59**	**0.02**	**168.10**	**457**	**<0.001**
USDT	**2.18**	**0.02**	**134.60**	**457**	**<0.001**
ESL	**2.39**	**0.02**	**144.66**	**457**	**<0.001**
UAI	**2.24**	**0.02**	**137.87**	**457**	**<0.001**

Table D.2 Unconditional Means Model Dimension-Specific Random Effects Estimates

Dimension	Standard Deviation	Variance Component	Degrees of freedom	χ^2	p-value
CERR					
Level 1: e_{ti}	0.53	0.28			
Level 2: r_{0i}	**0.30**	**0.09**	**457**	**1545.03**	**<0.001**
ECL					
Level 1: e_{ti}	0.56	0.32			
Level 2: r_{0i}	**0.29**	**0.09**	**457**	**1359.45**	**<0.001**
MCP					
Level 1: e_{ti}	0.54	0.29			
Level 2: r_{0i}	**0.29**	**0.08**	**457**	**1393.73**	**<0.001**
MSB					
Level 1: e_{ti}	0.49	0.24			
Level 2: r_{0i}	**0.30**	**0.09**	**457**	**1716.49**	**<0.001**
CS					
Level 1: e_{ti}	0.54	0.29			
Level 2: r_{0i}	**0.25**	**0.06**	**457**	**1186.74**	**<0.001**
USDT					
Level 1: e_{ti}	0.60	0.36			
Level 2: r_{0i}	**0.26**	**0.07**	**457**	**1080.29**	**<0.001**
ESL					
Level 1: e_{ti}	0.56	0.32			
Level 2: r_{0i}	**0.28**	**0.08**	**457**	**1270.36**	**<0.001**
UAI					
Level 1: e_{ti}	0.60	0.36			
Level 2: r_{0i}	**0.26**	**0.07**	**457**	**1104.59**	**<0.001**

Table D.3 Unconditional Growth Model Dimension-Specific Parameter Estimates

Fixed Effects	Coefficient	Standared Error	$t-ratio$	Approx. d.f.	p-value
CERR					
Intercept: β_{00}	**2.68**	**0.02**	**114.44**	**457**	**<0.001**
Slope: β_{10}	-0.0004	0.0010	-0.48	457	0.635
ECL					
Intercept: β_{00}	**2.47**	**0.02**	**103.49**	**457**	**<0.001**
Slope: β_{10}	-0.0002	0.0008	-0.20	457	0.843
MCP					
Intercept: β_{00}	**2.63**	**0.02**	**114.17**	**457**	**<0.001**
Slope: β_{10}	0.0008	0.0007	1.11	457	0.269
MSB					
Intercept: β_{00}	**2.75**	**0.02**	**113.04**	**457**	**<0.001**
Slope: β_{10}	-0.0004	0.0008	-0.54	457	0.588
CS					
Intercept: β_{00}	**2.60**	**0.023**	**114.94**	**457**	**<0.001**
Slope: β_{10}	-0.0008	0.0008	-0.10	457	0.319
USDT					
Intercept: β_{00}	**2.18**	**0.024**	**90.74**	**457**	**<0.001**
Slope: β_{10}	-0.0004	0.0009	-0.46	457	0.647
ESL					
Intercept: β_{00}	**2.36**	**0.02**	**100.05**	**457**	**<0.001**
Slope: β_{10}	0.0014	0.0008	1.60	457	0.109
UAI					
Intercept: β_{00}	**2.23**	**0.02**	**90.08**	**457**	**<0.001**
Slope: β_{10}	0.0007	0.0008	0.81	457	0.419

Table D.4 Novice Model Dimension-Specific Variance Components

Random Effects	Standard Deviation	Variance Component	d.f.	χ^2	p-value
CERR					
Level 1: e_{ti}	0.52	0.27			
Level 2					
Intercept: r_{0i}	**0.31**	**0.10**	**436**	**687.09**	**<0.001**
Slope: r_{1i}	**0.006**	**0.00003**	**436**	**503.44**	**0.014**
ECL					
Level 1: e_{ti}	0.56	0.31			
Level 2					
Intercept: r_{0i}	**0.29**	**0.09**	**436**	**650.16**	**<0.001**
Slope: r_{1i}	**0.007**	**0.00005**	**436**	**542.67**	**<0.001**
MCP					
Level 1: e_{ti}	0.54	0.29			
Level 2					
Intercept: r_{0i}	**0.29**	**0.08**	**436**	**624.03**	**<0.001**
Slope: r_{1i}	0.002	0.00000	436	428.23	>0.500
MSB					
Level 1: e_{ti}	0.48	0.23			
Level 2					
Intercept: r_{0i}	**0.37**	**0.14**	**436**	**842.07**	**<0.001**
Slope: r_{1i}	**0.008**	**0.00007**	**436**	**581.42**	**<0.001**
CS					
Level 1: e_{ti}	0.53	0.29			
Level 2					
Intercept: r_{0i}	**0.27**	**0.07**	**436**	**609.97**	**<0.001**
Slope: r_{1i}	**0.006**	**0.00004**	**436**	**498.41**	**0.020**
USDT					
Level 1: e_{ti}	0.59	0.34			
Level 2					
Intercept: r_{0i}	**0.27**	**0.07**	**436**	**580.39**	**<0.001**
Slope: r_{1i}	**0.007**	**0.00005**	**436**	**501.79**	**0.016**
ESL					
Level 1: e_{ti}	0.56	0.31			
Level 2					
Intercept: r_{0i}	**0.28**	**0.08**	**436**	**609.86**	**<0.001**
Slope: r_{1i}	**0.007**	**0.00005**	**436**	**514.56**	**0.006**
UAI					
Level 1: e_{ti}	0.59	0.35			
Level 2					
Intercept: r_{0i}	**0.29**	**0.09**	**436**	**596.49**	**<0.001**
Slope: r_{1i}	0.005	0.00002	436	454.82	0.257

Table D.5 Novice Model Parameter Estimates—CERR

Fixed Effects	Coefficient	Standared Error	$t - ratio$	Approx. d.f.	p-value
Intercept (Score Beginning Y1)					
Experienced: β_{00}	**2.74**	**0.03**	**11.65**	**456**	**<0.001**
Δ Novice: β_{01}	-0.0009	0.0008	-1.04	456	0.299
Slope (Growth per week)					
Experienced: β_{10}	**-0.32**	**0.06**	**-4.99**	**456**	**<0.001**
Δ Novice: β_{11}	0.003	0.0022	1.33	456	0.185
Random Effects	Standard Deviation	Variance Component	d.f.	χ^2	p-value
Level 1: e_{ti}	0.522	0.27			
Level 2					
Intercept: r_{0i}	**0.29**	**0.08**	**648.72**	**435**	**<0.001**
Slope: r_{1i}	**0.006**	**0.00003**	**501.03**	**435**	**0.015**

Table D.6 Novice Model Estimates—ECL

Fixed Effects	Coefficient	Standared Error	$t - ratio$	Approx. d.f.	p-value
Intercept (Score Beginning Y1)					
Experienced: β_{00}	**2.52**	**0.03**	**95.342**	**456**	**<0.001**
Δ Novice: β_{01}	-0.00043	0.0010	-0.444	456	0.657
Slope (Growth per week)					
Experienced: β_{10}	**-0.24**	**0.06**	**-4.307**	**456**	**<0.001**
Δ Novice: β_{11}	0.002	0.002	0.778	456	0.437
Random Effects	Standard Deviation	Variance Component	d.f.	χ^2	p-value
Level 1: e_{ti}	0.56	0.31			
Level 2					
Intercept: r_{0i}	**0.28**	**0.08**	**628.95**	**435**	**<0.001**
Slope: r_{1i}	**0.007**	**0.00005**	**541.63**	**435**	**<0.001**

Table D.7 Novice Model Estimates—MCP

Fixed Effects	Coefficient	Standared Error	$t-ratio$	Approx. d.f.	p-value
Intercept (Score Beginning Y1)					
Experienced: β_{00}	**2.66**	**0.02**	**109.21**	**456**	**<0.001**
Δ Novice: β_{01}	**-0.23**	**0.06**	**-3.592**	**456**	**<0.001**
Slope (Growth per week)					
Experienced: β_{10}	0.0008	0.0008	1.017	456	0.310
Δ Novice: β_{11}	0.0003	0.0020	0.155	456	0.877
Random Effects	Standard Deviation	Variance Component	d.f.	χ^2	p-value
Level 1: e_{ti}	0.54	0.29			
Level 2					
Intercept: r_{0i}	**0.27**	**0.07**	**605.06**	**435**	**<0.001**
Slope: r_{1i}	0.002	0.00000	427.94	435	>0.500

Table D.8 Novice Model Parameter Estimates—MSB

Fixed Effects	Coefficient	Standared Error	$t-ratio$	Approx. d.f.	p-value
Intercept (Score Beginning Y1)					
Experienced: β_{00}	**2.80**	**0.03**	**113.537**	**456**	**<0.001**
Δ Novice: β_{01}	**-0.35**	**0.07**	**-4.864**	**456**	**<0.001**
Slope (Growth per week)					
Experienced: β_{10}	-0.0011	0.0008	-1.368	456	0.172
Δ Novice: β_{11}	0.004	0.0023	1.878	456	0.061
Random Effects	Standard Deviation	Variance Component	d.f.	χ^2	p-value
Level 1: e_{ti}	0.48	0.23			
Level 2					
Intercept: r_{0i}	**0.34**	**0.12**	**790.79**	**435**	**<0.001**
Slope: r_{1i}	**0.00797**	**0.00006**	**576.05**	**435**	**<0.001**

Table D.9 Novice Model Parameter Estimates—CS

Fixed Effects	Coefficient	Standared Error	$t-ratio$	Approx. d.f.	p-value
Intercept (Score Beginning Y1)					
Experienced: β_{00}	**2.64**	**0.02**	**110.35**	**456**	**<0.001**
Δ Novice: β_{01}	**-0.24**	**0.06**	**-3.78**	**456**	**<0.001**
Slope (Growth per week)					
Experienced: β_{10}	-0.001	0.0009	-1.22	456	0.222
Δ Novice: β_{11}	0.002	0.002	0.805	456	0.421
Random Effects	Standard Deviation	Variance Component	d.f.	χ^2	p-value
Level 1: e_{ti}	0.54	0.29			
Level 2					
Intercept: r_{0i}	**0.26**	**0.07**	**592.77**	**435**	**<0.001**
Slope: r_{1i}	**0.006**	**0.00004**	**498.03**	**435**	**0.019**

Table D.10 Novice Model Parameter Estimates—USDT

Fixed Effects	Coefficient	Standared Error	$t-ratio$	Approx. d.f.	p-value
Intercept (Score Beginning Y1)					
Experienced: β_{00}	**2.21**	**0.03**	**83.75**	**456**	**<0.001**
Δ Novice: β_{01}	**-0.16**	**0.06**	**-2.640**	**456**	**0.009**
Slope (Growth per week)					
Experienced: β_{10}	-0.0004	0.001	-0.446	456	0.656
Δ Novice: β_{11}	0.0004	0.0022	0.175	456	0.861
Random Effects	Standard Deviation	Variance Component	d.f.	χ^2	p-value
Level 1: e_{ti}	0.59	0.34			
Level 2					
Intercept: r_{0i}	**0.26**	**0.07**	**571.70**	**435**	**<0.001**
Slope: r_{1i}	**0.0070**	**0.00005**	**501.53**	**435**	**0.015**

Table D.11 Novice Model Parameter Estimates—ESL

Fixed Effects	Coefficient	Standared Error	$t-ratio$	Approx. d.f.	p-value
Intercept (Score Beginning Y1)					
Experienced: β_{00}	**2.41**	**0.03**	**93.17**	**456**	**<0.001**
Δ Novice: β_{01}	**-0.25**	**0.06**	**-4.307**	**456**	**<0.001**
Slope (Growth per week)					
Experienced: β_{10}	0.0009	0.0009	0.925	456	0.356
Δ Novice: β_{11}	0.0027	0.0021	1.325	456	0.186
Random Effects	Standard Deviation	Variance Component	d.f.	χ^2	p-value
Level 1: e_{ti}	0.556	0.3090			
Level 2					
Intercept: r_{0i}	**0.27**	**0.07**	**595.27**	**435**	**<0.001**
Slope: r_{1i}	**0.007**	**0.00005**	**512.22**	**435**	**0.006**

Table D.12 Novice Model Parameter Estimates—UAI

Fixed Effects	Coefficient	Standared Error	$t-ratio$	Approx. d.f.	p-value
Intercept (Score Beginning Y1)					
Experienced: β_{00}	**2.25**	**0.03**	**81.429**	**456**	**<0.001**
Δ Novice: β_{01}	**-0.13**	**0.06**	**-2.172**	**456**	**0.030**
Slope (Growth per week)					
Experienced: β_{10}	0.0008	0.001	0.845	456	0.398
Δ Novice: β_{11}	-0.00045	0.0019	-0.231	456	0.817
Random Effects	Standard Deviation	Variance Component	d.f.	χ^2	p-value
Level 1: e_{ti}	0.591	0.349			
Level 2					
Intercept: r_{0i}	**0.29**	**0.08**	**591.11**	**435**	**<0.001**
Slope: r_{1i}	0.005	0.00002	454.62	435	0.249

E: Assumptions of HLM

This appendix provides a detailed discussion regarding the assumptions of HLM applied to the novice model as well as investigations of the extent to which we are convinced the assumptions are met.

The assumptions of HLM are as follows:

1) Each e_{ti} is independent and normally distributed with a mean of 0 and constant variance for every level-1 unit within each level-2 unit, or occasion within teacher [i.e. e_{ti}~iid $N(0, \sigma^2)$].
2) The level-1 predictor ($Week_{ti}$) is independent of e_{ti}, or there is no covariance between the error terms and the predictor variable.
3) The vector of random effects at level-2 are multivariate normal, each with a mean of 0 and some constant variance. The random effects are independent among the level-2 units. [i.e. $r_{qi} = (r_{0i}, \ldots . r_{qi})'$~ iid $N(0, \boldsymbol{T})$]
4) The level-2 predictor ($Novice_i$) is independent of every level-2 random effect. In other words, the covariance of $Novice_i$ and the level-2 random effects is zero.
5) The random effects at levels-1 and -2 are all independent of one another.
6) The predictors at each level are not correlated with the random effects at other levels. (Adapted from Raudenbush & Bryk, 2002, p. 255).

These assumptions manifest themselves as such in the novice model:

1) Conditional on the $Week_{ti}$variable, the within-teacher errors are normal and independent with a mean of 0 for each teacher and equal variances across teachers.

2) Whatever occasion-level predictors of FFT scores are excluded from the model and thereby included in the error term e_{ti} are independent of the $Week_{ti}$variable included in the model.
3) The residual teacher effects (r_{0i}, r_{1i}) are assumed to have normal distributions with constant variances and covariance.
4) The effect of whatever teacher predictors are excluded from the model for the intercepts and slopes are independent of the $Novice_i$ variable.
5) The error at level-1, e_{ti} is independent of the residual teacher effects, r_{0i}, r_{1i}.
6) Whatever occasion-level predictors that are excluded from the level-1 model and thereby relegated to the error term e_{ti} are independent of the level-2 predictor, $Novice_i$. In addition, whatever teacher-level predictors are excluded from the model and thereby relegated to the level-2 random effects, are uncorrelated with the occasion-level predictors.

In order to assess Assumption 1, consider both a histogram of the level-1 residuals as well as a scatter plot of these residuals by a random Teacher-Occasion ID in Figure E.1.

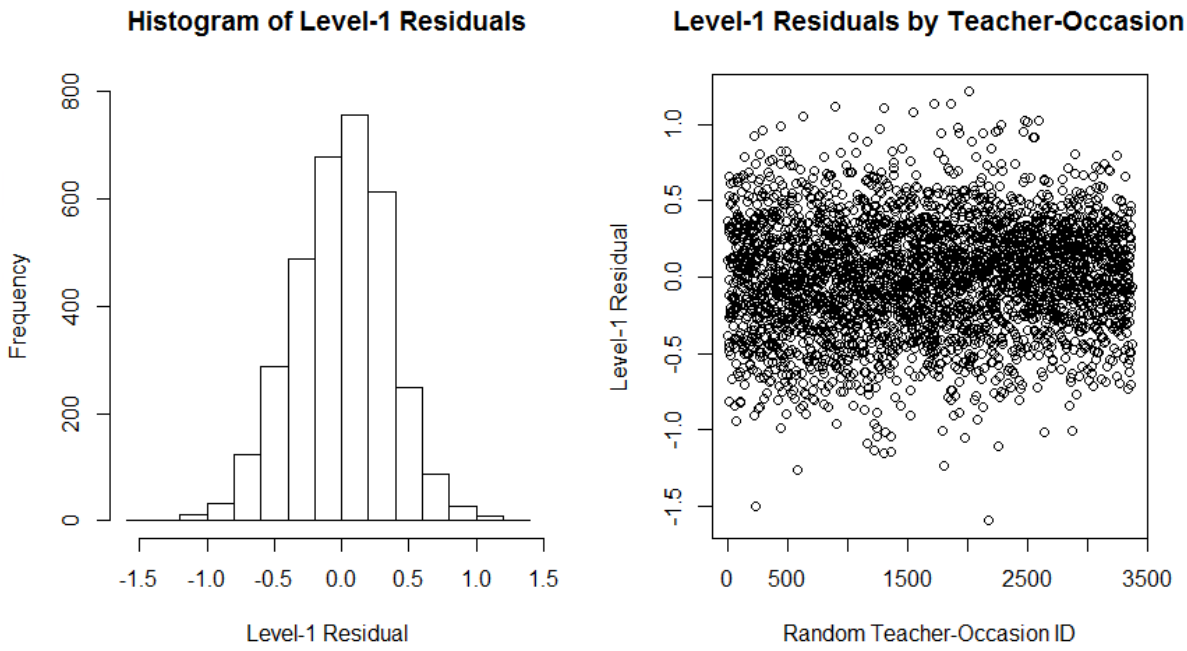


Figure E.1 Level-1 residuals

Figure E.1 illustrates the level-1 residuals from the novice model with mean FFT-dimension score as the outcome of interest. The left panel of Figure E.1 shows a normal distribution of these residuals, and the right panel indicates homogeneous variance across teacher-occasions. This data suggests that the first assumption of HLM is not violated. However, it is important to remember that these distributions are for residuals and not errors. We use the residuals to estimate errors and assume that they adequately indicate what is happening with the actual errors. In addition to the distributions illustrated in Figure E.1, the covariance between the level-1 residuals and the $Novice_i$ control variable is 0. This suggests that Assumption 2 is not violated in the novice model either.

In order for Assumption 3 to be met, all of the level-2 residuals need to be normally distributed and have constant variances and covariances. Evidence to this point are provided in Figures E.2 – E.3.

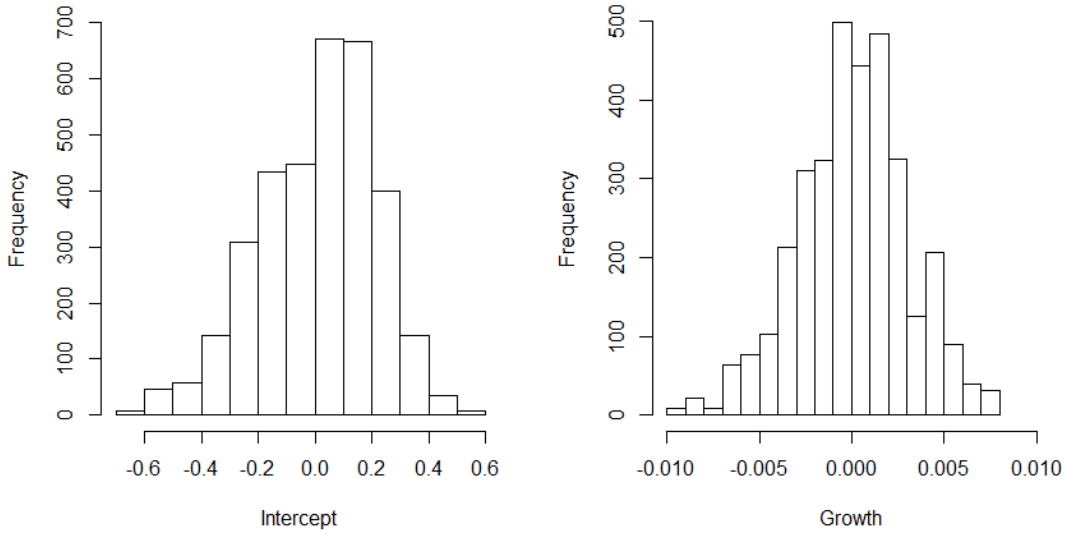


Figure E.2 Histograms of Level-2 residuals

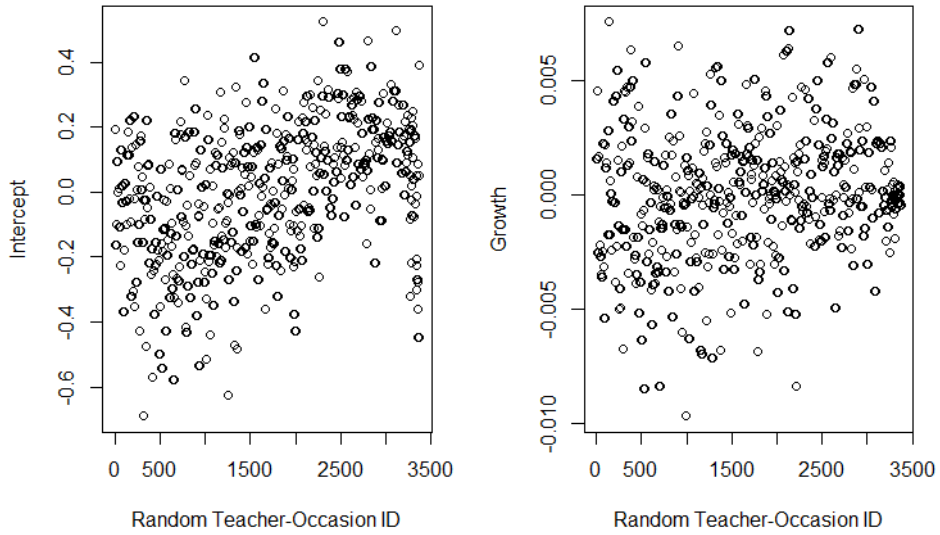


Figure E.3 Scatterplots of Level-2 residuals across Teacher IDs

Figure E.2 indicates that both of the level-2 residual distributions are relatively normal with no large, detectable skew in either direction. Figure E.3 shows that the variance of the residuals for the slopes and intercepts are not constant across all teachers. Specifically, FFT scores for teachers on the right side of the scatterplots appear to have smaller variances than others as indicated by the narrowing of the cloud of points on the right side. Despite this fact,

these violations to Assumption 3 are relatively modest. The violations may make us somewhat concerned regarding the efficiency in the fixed effects estimates as well as incorrect shrinkage of the level-1 coefficients. However, this does not bias the level-2 estimates.

In order to assess Assumptions 4 – 6, consider the table of covariances (expressed as correlations) between the level-1 residuals, the level-2 residuals, and the predictor variables in Table E.1. Evidence for Assumption 4 comes from the relationship between the level-2 residuals and the $Novice_i$ variable. The results presented in Table E.1 suggest no evidence that Assumption 4 is violated based on our information from the model residuals. More specifically, we see this correlation to be about zero. In order to meet Assumption 5, the covariance between the level-1 residual and the levels-2 residuals must also be zero. It is clear from the middle section of the first column in Table E.1 that this is not the case. Specifically, the correlation between the level-1 residual and the level-2 residuals ranges from 0.04 to 0.18. Finally, in order for Assumption 6 to be met, the correlation between each of the level-1 predictors must not be related to the level-2 residuals. The bottom section of Table E.1 indicates that there is little evidence to suggest this assumption is violated.

Table E.1 Correlation Matrix: Level-2 Residuals, Level-1 Residuals, and Predictors

Residuals	Level-1 Residual	Level-2 Residuals		Predictors	
	e_{ti}	Intercept: r_{0i}	Slope: r_{1i}	Novice	Week No Summer
Level-1:					
e_{ti}	1.00				
Level-2					
Intercept: r_{0i}	0.18	1.00			
Slope: r_{1i}	0.04	0.09	1.00		
Predictors					
Novice	< -0.01	< 0.01	< 0.01	1.00	
Week No Summer	< 0.01	-0.01	0.01	0.02	1.00

The most glaring violation of the HLM assumptions is Assumption 5, the independence of the residuals across levels. Consider an observation that occurs just days before a major holiday break. It is very likely the case that an observation on this day will affect the teacher as well as that specific occasion and the dimension-level scores of the day. Perhaps the teacher is distracted about catching a flight out to see family that evening. In addition, the lesson on that particular occasion might be misrepresentative of the typical sort of lesson that teacher might deliver. Finally, the student and teacher behaviors that inform the dimension-level scores might be significantly altered due to the anticipation of the long-awaited break. Each of these elements will contribute to the random effect at each level of the HLM, but they will all surely be related to one another.

Violating the assumption of the independence of random effects across levels does not necessarily bias the level-2 coefficients, but it can have negative effects on the estimated standard errors and the related inferential statistics. The results from the current study indicate that differences in growth rates for novice teachers are non-significant. The likelihood that the error structure includes covariance across levels suggests that these differences are almost certainly non-significant. The p-values for the parameters associated with experienced teachers are quite a bit lower ($\leq$.005 for all parameters), providing relatively stronger levels of confidence in the parameter estimates. However, the violations to the assumptions regarding the error structure of the model might lessen our confidence in the standard errors of the estimates. When differences between groups are small, it is important for the standard errors to be as small as possible in order to have the best chance at detecting such differences. Thus, violation of this assumption is also a limitation of the current study.