EDUCATION AND LABOR

Validation Study of the TNTP Core Teaching Rubric

Andrew McEachin, Jonathan Schweig, Rachel Perera, Isaac M. Opper

Sponsored by TNTP

For more information on this publication, visit www.rand.org/t/RR2623

Published by the RAND Corporation, Santa Monica, Calif.

Preface

TNTP (formerly The New Teacher Project) commissioned a multiyear validation study of the the TNTP Core Teaching Rubric (TNTP Core), an observation-based measure of teachers' Common Core State Standards–aligned instruction. A distinctive feature of TNTP Core is that ratings are based on student behaviors rather than teacher behaviors. TNTP asked the RAND Corporation to conduct an independent study focused on two key tasks: (1) collect evidence of the extent to which the content expertise of raters may influence TNTP Core scoring, and (2) collect evidence on the extent to which TNTP Core scores relate to other measures of instructional quality.

The results of the report should be useful to practitioners interested in using TNTP Core, or other observable rubrics, as a tool to measure teacher practice. We highlight a number of challenges that practitioners and policymakers may face when using observational rubrics to make broad inferences about teacher quality. The results also address a number of technical and measurement issues related to the design and implementation of observational rubrics.

This study was undertaken by RAND Education and Labor, a division of the RAND Corporation that conducts research on early childhood through postsecondary education programs, workforce development, and programs and policies affecting workers, entrepreneurship, financial literacy, and decisionmaking. This study was sponsored by TNTP with the support of the Charles and Lynn Schusterman Family Foundation. TNTP is a nonprofit organization that works within every level of public school systems—including school districts, state departments of education, and charter-school networks—to improve student educational outcomes and reduce educational inequality by working with high-need schools to attract, train, and retain teachers.

More information about RAND can be found at www.rand.org. Questions about this report should be directed to Andrew McEachin at mceachin@rand.org, and questions about RAND Education and Labor should be directed to educationandlabor@rand.org.

Contents

Figures

Tables

Summary

Nearly every public school teacher in the United States receives some form of feedback based on classroom observations. Practitioners and policymakers use scores derived from observation protocols both as indicators of teaching quality and as input for administrators, coaches, and other instructional leaders to provide teachers with detailed, continuous feedback for professional development purposes.

One recently developed observation protocol is the TNTP Core Teaching Rubric (TNTP Core) (TNTP, 2014).[1] Unlike other observation rubrics that are scored based on teacher actions, TNTP Core focuses on student behavior. The rubric contains four domains (Culture of Learning, Essential Content, Academic Ownership, and Demonstration of Learning) that measure the extent to which students are engaging in or demonstrating a specific behavior, with each domain scored on a five-point scale. TNTP Core was designed to efficiently describe high-quality, Common Core State Standards–aligned instructional practice, and it is intended to help teachers and administrators focus on actionable feedback and to promote teachers' professional development within the context of a multidimensional evaluation system. Given the potential for TNTP Core to promote positive teacher development and high-quality instruction that supports the implementation of the Common Core State Standards, it is important to collect evidence that scores from the rubric are accurate, are not influenced by the assignment of observers to teachers, and are positively associated with other measures of teacher performance.

TNTP asked the RAND Corporation to conduct an independent study focused on two key tasks: (1) collect evidence of the extent to which the content expertise of raters may influence TNTP Core scoring, and (2) collect evidence on the extent to which TNTP Core scores relate to other measures of instructional quality.

Overview of the Study Design

Volunteer teachers of mathematics and English language arts (ELA) from more than 20 districts and charter school networks across eight states contributed to this study by having their classroom instruction videotaped three times during both the 2015–2016 and 2016–2017 school years. This sample of volunteer teachers was complemented with a sample of approximately 100 4th- and 5th-grade teachers from the Measures of Effective Teaching (MET) study for whom videotaped instruction was available (Measures of Effective Teaching Longitudinal Database, undated). The videos from both samples were scored on TNTP Core by trained raters. In both settings, the taped lessons were used only for research purposes, and teachers' TNTP Core scores

[1] TNTP (formerly The New teacher Project) is a nonprofit organization that works within every level of public school systems—including school districts, state departments of education, and charter-school networks—to improve student educational outcomes and reduce educational inequality by working with high-need schools to attract, train, and retain teachers.

had no stakes attached. We use teachers' math and ELA value-added scores as the alternative measure of instructional quality. Value-added scores capture teachers' contributions to students' math and ELA achievement, excluding factors outside teachers' control. The value-added scores for this study were obtained or calculated for a subset of teachers.

While this study relies on a rigorous research design, there are a few limitations to keep in mind. First, results from validity studies are inherently contextual, and our findings are based on a small sample of volunteer teachers. Second, our study focuses on three aspects of validity: scoring consistency and bias, generalization of rater scores, and extrapolation of rater scores to other measures of teacher quality. Our research design does not cover all aspects of validity important to the design and implementation of TNTP Core.

Findings

Relationships between teachers' TNTP Core scores and student achievement gains are modest and vary by subject area.

Our results suggest that TNTP Core scores correlate with teacher value-added scores, though the strength of these relationships differs by subject area and by TNTP Core domain. We find correlations for math between 0 and 0.12 and correlations for ELA between 0.03 and 0.21. A correlation of 0.2 suggests that TNTP Core scores explain just 4 percent of teachers' value-added score ranking.

Observers often disagree in their ratings of instructional practice based on TNTP Core, and these disagreements may be related to their content expertise.

Across all four domains and subjects, on average, raters struggled to agree on their judgments about the quality of a lesson. Overall, agreement rates were no better than what would be expected by chance. Raters also varied in the severity of lesson ratings in math, and this was in part based on content knowledge. Math specialist raters in particular were more likely to rate a lesson lower than their generalist peers.

There is a considerable amount of uncertainty about the extent to which TNTP Core scores represent teachers' overall instructional practices.

Disagreements among raters introduce uncertainty into TNTP Core scores and weaken claims about teacher practices that can be made on the basis of those scores. However, raters are not the only source of uncertainty in TNTP Core scores. There is also uncertainty that is introduced based on which specific lessons are observed and rated. To make general claims about a teacher's instructional quality from TNTP Core scores, one would need many observations of a given teacher conducted by many raters throughout the year. Our results suggest that even if teachers were observed four times per year, each time by two raters, there would still be considerable uncertainty about the quality of teachers' instructional practice. This uncertainty has at least two consequences. First, it attenuates (i.e., weakens) relationships between TNTP Core scores and other indicators of teacher quality, making it difficult to assess the extent to which the aspects of instruction measured by TNTP Core are related to student outcomes. Second, it diminishes the usefulness of feedback that can be given based on TNTP Core scores. If TNTP

Core scores do not accurately represent the practices of teachers overall, then their utility in diagnosing and developing instructional practices is compromised.

Conclusion and Recommendations

Taken together, these results have several implications for schools, school districts, and charter networks intending to use TNTP Core to provide teachers with feedback on their performance. We conclude with four issues that merit consideration as a part of implementation to maximize the usefulness of TNTP Core.

Set High Standards for Rater Certification

Currently, TNTP uses a 75/50 threshold for rater certification. In order to be considered qualified to use TNTP Core to rate instruction, across five lessons trainees must reach 75 percent agreement-within-one and 50 percent exact agreement with a master rater.[2] We suggest raising the criteria scores for rater certification as one potential method to increase rater agreement. While meeting certification requirements is no guarantee that, in practice, raters will rate consistently or accurately, it is possible that higher certification standards can promote increased rater agreement. At a minimum, criterion scores should account for the possibility of chance agreement. For this reason, adopters may even want to consider using 90 percent agreement-within-one and 70 percent exact agreement as criteria scores for rater certification, in line with other certification processes (Kane & Staiger, 2012; Bell et al., 2012).

Build Systems to Monitor the Score Quality Under Operational Conditions

Many factors influence the quality of observation scores, including the context in which scores are collected, the extent to which teachers vary in their instructional practices, and the extent to which raters apply scoring rules consistently and accurately. While initial certification and training are important components in ensuring that scoring rules are applied with fidelity, research has shown that raters tend to change their approach to scoring over time—a phenomenon known as rater drift. Rater drift begins to occur almost immediately after training and continues over time (Casabiana, Lockwood, & McCaffrey, 2014). It may be helpful for schools and school districts to embed more frequent post-certification calibration and validation exercises during a rating period (Bergin et al., 2017). Calibration could involve having all participating raters score the same (master-rated) lesson video and discuss score inconsistencies. Validation could involve monitoring raters' scores by having a master rater score a subset of the lesson videos. It may be helpful to conduct these activities as often as weekly (Bell et al., 2012).

High-Quality Evaluation and Feedback Require Many Observers with Different Backgrounds to Rate Many Lessons

Consistent with other research (Kane & Staiger, 2012), we found that assertions about a teacher's overall instructional practices based on TNTP Core scores would be valid only if they were

[2] *Agreement-within-one* refers to two raters assigning scores within one point of each other. For example, if one rater assigned a score of 3 and another rater assigned a score of 4, this would constitute agreement-within-one.

based on many observations of a given teacher conducted by many raters throughout the year. There is likely to be considerable uncertainty about a teacher's overall instructional practices even when assertions about instructional practice are based on four observations conducted by two raters in a given school year. Our findings suggest that scores also depend on the specific rater, because content expertise can influence the severity or leniency of ratings. This suggests not only that many raters be used, but also that it is important to ensure that teachers are observed by both content experts and content generalists. It may also be helpful to administer a content knowledge assessment (in this study, we used the Content Knowledge for Teaching [CKT]) to all prospective raters, in order to gain some insight into content expertise prior to engaging in classroom observations.

Consider Collecting Additional Evidence of Validity

Reliability and validity are not static properties of observation protocols; they are better thought of as processes—they depend, for example, on who is observed and the conditions under which scores are collected. Scores collected from researchers or external observers may have different properties than scores collected from principals or peers. Additionally, validity is closely tied to how scores are intended to be used. Scores could potentially be used in a wide range of ways: to promote conversations within grade-level or content-area teams, to guide or inform professional development plans, or to guide or inform decisions during performance evaluations. Different kinds of evidence would be necessary to support these intended uses, and uses that are tied to consequences require a strong evidence base and a lower tolerance for uncertainty. As schools, school districts, and charter school networks outline specific uses for TNTP Core, they should consider collecting other sources of evidence that support claims about the quality of teacher practice—including evidence based on the content of the protocol (Does it adequately measure the intended dimensions?), evidence based on outcomes and consequences, and evidence based on response processes (how raters are interpreting and scoring the domains). If TNTP Core is used as one measure in a multiple-measure system, separate validity evidence should be collected supporting inferences based on multiple measures.

Acknowledgments

The authors are grateful to Courtney Bell and Daniel McCaffrey for providing feedback on various aspects of the study and administering and scoring the Content Knowledge for Teaching (CKT). This report benefited substantively from feedback from Kata Mihaly and Fatih Unlu at the RAND Corporation; from Jose Felipe Martínez at the University of California, Los Angeles; and from Erin Grogan, Amy Hammerle, Cassandra Coddington, and their colleagues at TNTP. James Torr provided expert editing. Any flaws or errors that remain are the sole responsibility of the authors.

Abbreviations

CCSS	Common Core State Standards
CKT	Content Knowledge for Teaching
TNTP Core	the TNTP Core Teaching Rubric
ELA	English language arts
G-theory	generalizability theory
MET	Measures of Effective Teaching
SD	standard deviation
SE	standard error
VA	value-added score
VAM	value-added model

1. Introduction

Teacher observation is widely used to provide information about instructional practice to researchers, practitioners, and policymakers (see, for example, Cohen & Goldhaber, 2016; Martínez, Borko, & Stecher, 2011; Pianta & Hamre 2009; Rowan & Correnti, 2009). Scores derived from observation protocols can be used as summative indicators of teaching quality and can enable administrators, coaches, and other instructional leaders to provide teachers with detailed, continuous feedback for professional development purposes (Goldring et al., 2015; Hsieh et al., 2009; Pianta & Hamre, 2009; Mihaly et al., 2018; Taylor & Tyler, 2012). Nearly every state and local school district in the United States requires that teachers receive some form of feedback based on classroom observations (Doherty & Jacobs, 2013), and as many as 95 percent of public school teachers are evaluated based on formal classroom observation (Cohen & Goldhaber, 2016).

Advocates believe that teacher observation protocols clearly and transparently define high-quality instruction and that ratings based on observation are consistent with beliefs about best practices (Goldring et al., 2015; Goe, Bell, & Little, 2008). However, perhaps in an effort to adequately describe a practice as complex as instruction, many observation instruments are also long and complex, containing dozens of indicators, elements, and domains. The complexity of these instruments presents significant challenges to successful implementation. First, using long, complicated instruments creates a substantial administrative burden. Principals often invest substantial amounts of time in the observation process, and, as a result, administrators in some states have noted that they are not able to provide thorough evaluations or meaningful feedback to teachers (U.S. Government Accountability Office, 2013; Reform Support Network, 2013). Second, some research has suggested that even when there is adequate time to thoroughly attend to the observation process, it is difficult for observers to keep track of multiple indicators at the same time, which compromises the overall quality and accuracy of ratings (Kane & Staiger, 2012).

At the same time, observation-based measures of instructional practice are invariably linked to changes to state content standards. Since 2011, 45 U.S. states and the District of Columbia have formally adopted the Common Core State Standards (CCSS) for mathematics and for English language arts (ELA).[1] These standards represent a substantial departure from most previous state standards and require teachers to shift their instructional practice for successful implementation. For example, CCSS places greater demands on teachers to embed cognitively challenging instructional tasks in classroom work and utilize more-sophisticated pedagogical content knowledge to support students' work on those tasks (Ball & Forzani, 2011). As instructional needs and goals change and teachers respond to the instructional demands of curricula aligned with these new standards, the theories of quality practice on which observational rubrics are built

[1] Although many states have made some subsequent changes to those standards after initial adoption, most states have retained key aspects of the original CCSS (Achieve, 2017; Norton, Ash, & Ballinger, 2017; Korn, Gamboa, & Polikoff, 2016).

may not align with teachers' new instructional demands (Cohen & Goldhaber, 2016). This is perhaps why policymakers have recommended that states encourage and prioritize the representation of CCSS instructional shifts in teacher evaluation (Wiener, 2013).

One recently developed observation protocol that was designed to address these concerns is the TNTP Core Teaching Rubric (TNTP Core), developed by TNTP (formerly The New Teacher Project) in 2014. TNTP is a nonprofit organization that works within every level of public school systems—including school districts, state departments of education, and charter-school networks—to improve student educational outcomes and reduce educational inequality by working with high-need schools to attract, train, and retain teachers. TNTP designed TNTP Core to efficiently describe high-quality, CCSS-aligned instructional practice, and TNTP Core is intended to help teachers and administrators focus on actionable feedback and to promote the professional development within the context of a multidimensional evaluation system (TNTP, 2014).

Study Purpose and Research Questions

Given the potential for TNTP Core to promote effective teacher development and high-quality instruction that supports the implementation of CCSS, it is important to collect evidence that scores are accurate, are not influenced by the assignment of observers to teachers, and are positively associated with other measures of teacher performance. In other words, it is important to demonstrate that the scores support valid inferences about instructional practice (Kane, 2006; Bell et al., 2012).

The purpose of this report is to provide validity evidence to support the use of TNTP Core to provide teachers with feedback and to evaluate their performance. This report is based on data collected during the 2015–2016 and 2016–2017 school years from more than 20 school districts and charter networks in eight states, all of which implemented the CCSS or similar standards. These data were complemented with data from the Measures of Effective Teaching (MET) study sample (Measures of Effective Teaching Longitudinal Database, undated). The data for our study include video recordings of teacher practice that were scored by trained raters, as well as value-added scores (either obtained or calculated) for a subsample of participating teachers. Specifically, this study aims to address the following two research questions:

1. To what extent does content expertise of the rater influence scores on TNTP Core?
2. To what extent are teacher observation scores (overall and on each of the four rubric domains) valid predictors of the effectiveness of math and ELA teachers, as captured by their estimated contributions to student performance (or value-added scores)?

The objective of these research questions is to better understand the validity of TNTP Core scores. In this report, we first provide some evidence of the extent to which scoring rules are consistently applied and bias-free. We do this by examining the extent to which raters agree with one another in their ratings of the same teacher, and the extent to which rater background influences scores.

Second, we provide evidence of the extent to which scores can be extrapolated to represent instructional practice more broadly defined. We do this by examining the extent to which

observation scores correlate with measures of student performance—specifically, with teacher value-added scores that measure teacher effectiveness in math and ELA.

Although it was not a direct focus of the two research questions, we also appraise the reliability of TNTP Core scores. Reliability is a critical consideration to any validity study, because it provides information about the extent to which TNTP Core scores adequately characterize instructional practice by accounting for the various sources of error that are involved in assigning scores. Information about reliability is important for two reasons. First, teachers, principals, and other practitioners are often concerned that observation scores based on a sample of lessons may not be reflective of the quality of a teacher's practice overall. Information about reliability provides insight into the extent to which observed lessons and resulting scores reflect teacher practice broadly and are not limited to a specific time, lesson, or observer (Kane, 2006; Bell et al., 2012). Second, reliability is a necessary precondition for understanding whether TNTP Core scores are associated with student achievement gains. The presence of error can make it more difficult to understand the extent to which instructional practices are related to student outcomes. If observed relationships are not sensitive to the specific lessons that were scored, or to the specific observers that were assigned, this strengthens claims about the relationship between instructional practice and student performance.

Report Organization

This report is divided into seven chapters. Chapter Two provides an overview of the study design, including information about the sampling of teachers, and descriptions of TNTP Core, rater training, and the assignment of raters. Chapter Three covers our analytic methods. Chapter Four describes findings about the extent to which scores are consistent and bias-free, including whether rater background influences scores. Chapter Four also describes the extent to which generalized claims about instructional practices based on TNTP Core scores are warranted. Chapter Five describes findings about the extent to which teacher observation scores are valid predictors of student performance for math and ELA teachers. Chapter Six provides a summary and discussion of the results and closes with recommendations for practitioners and policymakers. Chapter Seven summarizes the main limitations of this study.

2. Design of Study

In this chapter, we describe the design of the study, including details about the sampling of teachers and lessons, the observation instrument, rater recruitment and training, and the assignment of raters to lessons. This study made use of two distinct samples of teachers, which we refer to as the *TNTP sample* and the *MET sample* throughout this report. Our analysis combines these samples, unless otherwise noted. Details about these samples are provided below, and for each sample we describe recruitment, training, and scoring processes separately.

Sampling of Teachers and Lessons

TNTP Sample

TNTP initiated teacher recruitment at the start of the 2015–2016 school year. Enrollment in the study was voluntary, and teachers were given $50 gift cards for participating in the study. TNTP followed a two-step process for recruiting study participants. First, TNTP approached a local education agency, school district, or charter school network to secure permission to recruit teachers. Once permission was granted, TNTP then directly recruited math and ELA teachers in elementary and middle school grades. During the 2015–2016 school year, TNTP was given permission to recruit teachers in two charter school networks (located in different states) with 23 total volunteer teachers. To increase the size of the study sample, TNTP recruited 97 more teachers in the 2016–2017 school year across 20 sites (consisting of both local school districts and charter school networks), some of which included teachers from TNTP's teaching fellow program.[1] These 20 sites were located across eight different states.

In both the 2015–2016 and 2016–2017 academic years, the non-fellow volunteer teachers had three lessons recorded. The time interval between observations was not standardized and varied considerably across teachers. The scheduling of the lessons was coordinated between the teacher and a local videographer in charge of collecting taped lessons in the district or charter network. The TNTP fellows were also observed (in-person) three times during the year and were given advance notice about the week (but not day) of the observation.[2]

To answer the second research question, we needed not only volunteer teachers with TNTP Core scores but also teachers with enough data to generate math and ELA value-added scores. Many of our research sites had a small number of teachers (e.g., fewer than five) and did not have previously estimated value-added scores (e.g., from a state accountability policy). Therefore, we

[1] More information about the TNTP Teaching Fellows program can be found here: http://tntpteachingfellows.org/

[2] We describe the entire sample of participating TNTP teachers here. For analyses of score consistency, rater bias, and score generalizability, we use a subsample of teachers from the overall TNTP sample. For analyses of relationships among TNTP Core scores and student achievement, we restricted our sample to locations that had at least five teachers in a given grade and subject with both videotaped lessons and value-added scores. More details about analytic subsamples are provided in Chapter Three.

had to narrow our sample to sites that either had enough teachers to estimate a value-added model (VAM) or provided an already estimated value-added score and correlate these value-added scores with teachers' TNTP Core scores. We were left with two subsamples that met this criteria. We refer to one of these subsamples as TNTP sample 1, which includes 25 ELA volunteer teachers and five math volunteer teachers across eight sites from a single state and the 2016–2017 school year. The second subsample, TNTP sample 2, is a large charter network operating on the East coast. This sample includes 17 ELA volunteer teachers and 15 math volunteer teachers from the 2015–2016 and 2016–2017 school years. We describe these samples and how they are used to answer our second research question in more detail in Chapter Three.

Measures of Effective Teaching Sample

From 2009 to 2012, the MET project (Measures of Effective Teaching Longitudinal Database, undated) collected extensive data on teacher practice, including student surveys, teacher surveys, lesson videos, and a wide range of achievement measures for a sample of more than 3,000 K–12 teachers in six large urban districts across the United States. Our goal with the MET data was to sample approximately 100 teachers who had at least four available lessons in the MET video data in both math and ELA. The MET sample consists of a randomly selected subset of 93 4th and 5th grade classrooms, from the 2009–2010 academic year, all drawn from a single district with lesson videos available in the MET project classroom video library. Although the broader MET study draws on data from six districts, we sampled from a single district to remove district effects and maximize our statistical power to answer our second research question. We randomly assigned these 93 teachers to either the math or ELA group. Furthermore, we randomly selected two lessons from each teacher in their assigned subject area.

The content of the video lessons differed slightly between the TNTP and MET samples. In the TNTP sample, the video recordings were able to capture both teacher and student actions. The location of the camera in the TNTP recordings captured audio from both the teacher and the students. In the MET videos, one can observe the teacher and his or her students, but only the teacher wore a microphone, so raters were able to hear only the teachers and those students who happened to be near the teacher. For the TNTP sample, TNTP used stated district policies to acquire student consent, which in most cases required passive consent (e.g., students were given the opportunity to opt out of the recorded lessons).

TNTP Core and Scoring Procedures

TNTP Core was designed to describe, assess, and provide actionable feedback on CCSS-aligned instructional practice while focusing more on student behavior than teacher actions (TNTP, 2014). TNTP Core describes instructional practice across four domains of performance: Culture of Learning, Essential Content, Academic Ownership, and Demonstration of Learning (see Table 2.1 for domain descriptors). All domains are rated on a five-point scale, from 1 (ineffective) to 5 (skillful). Observers assign ratings by making judgments about which combination of domain

descriptors most closely describes the observed instruction, based on a preponderance of evidence. Overall ratings for a domain are generated by averaging across observers and lessons.[3]

Table 2.1 Teacher Performance Areas Measured by TNTP Core

Domain	Descriptor
Culture of Learning	The extent to which all students are engaged in the work of the lesson from start to finish
Essential Content	The extent to which all students are engaged in content aligned to the appropriate standards for their subject and grade
Academic Ownership	The extent to which all students are responsible for doing the thinking in this classroom
Demonstration of Learning	The extent to which all students demonstrate that they are learning

NOTE: All domain scores range from 1 to 5. 1 = ineffective, 2 = minimally effective, 3 = developing, 4 = proficient, 5 = skillful.

Raters and Training

TNTP recruited 39 individuals as raters to score videos, and 34 passed the training requirements.[4] By design, these raters included both content generalists (N=24) and content area experts in math (N=7) and ELA (N=4). All raters were experienced educators with deep understanding of general pedagogy. To evaluate content area expertise, RAND subcontracted with ETS (Educational Testing Service) to administer the Content Knowledge for Teaching (CKT). The CKT is a subject specific assessment aimed at measuring whether teachers can apply specific content knowledge to their instruction and assessments of student learning. Four versions of the CKT were used: Math, grades 4 to 5; Math, grades 6 to 8; ELA, grades 4 to 6; and ELA, grades 7 to 9.

The generalist raters took all four assessments, and the content experts took the two assessments in their content area. Table 2.2 presents the average percentile scores for the CKT assessment by rater type. For the math assessments, math raters answered a higher percentage of the questions correctly (79 percent and 75 percent) than did generalists (65 and 62 percent). For the ELA assessments, the ELA raters answers the questions correctly at a more similar rate (76 percent and 73 percent) to generalists (79 and 69 percent).[5]

[3] In this report, we use the terms *rater* and *observer* interchangeably, and the terms *lesson* and *occasion* interchangeably.

[4] A small subset of the observation scores included in this report are based on live observations.

[5] The CKT was designed for the MET study, for which more than 2,000 teachers were administered the exam. In that sample, teachers averaged the following percentages correct: 52 for Math, grades 4 to 5; 62 for Math, grades 6 to 8; 66 for ELA, grades 4 to 6; and 65 for ELA, grades 7 to 9.

Table 2.2. Raters' Scores on ETS CKT Exam in Percentage Correct (N=35)

Rater	Math 4 to 5	Math 6 to 8	ELA 4 to 6	ELA 7 to 9
Generalist	65	62	79	69
Math	79	75		
ELA			76	73

All raters underwent 20 to 25 hours of virtual training prior to participating in rating activities. Training activities provided raters with opportunities to review content-embedded pedagogical practices, engage with the CCSS, and watch and score exemplar instructional videos using TNTP Core. At the conclusion of this training, all potential raters rated five lesson videos that had been master-scored, and they received feedback on the rating process (via webinar) after each scoring session. In order to pass training and be certified to rate videos during the study, raters had to reach predetermined benchmarks for accuracy. Specifically, across five lesson videos, raters needed to average agreement-within-one with master ratings 75 percent of the time, and exact agreement 50 percent of the time on each of the four scored domains.[6] We refer to this as the *75/50 threshold* throughout this report. Raters who passed the initial certification repeated the certification process three more times: in the summer of 2016, once during the 2016–2017 school year, and again prior to reviewing the MET lessons in the winter of 2018.

Assignment of Raters

TNTP Sample

An algorithm was developed by TNTP to randomly assign raters to videos. Every week, the lessons uploaded during the previous seven days were assigned randomly to be rated by both a generalist and a content area expert. Raters were randomly assigned based on availability, so some raters provided considerably more scores than other raters. Table 2.3 illustrates rater assignment based on a scenario with 10 teachers and 30 math lessons. A similar design would be used for ELA lessons.

[6] *Agreement-within-one* refers to two raters assigning scores within one point of each other. For example, if one rater assigned a score of 3 and another rater assigned a score of 4, this would constitute agreement-within-one.

Table 2.3. Illustrative Rater Assignment Table for Math Teachers: TNTP Sample

Teacher	Lesson 1		Lesson 2		Lesson 3	
1	GEN-006	MAT-003	GEN-001	MAT-004	GEN-007	MAT-004
2	GEN-006	MAT-004	GEN-006	MAT-002	GEN-004	MAT-001
3	GEN-001	MAT-002	GEN-005	MAT-004	GEN-007	MAT-004
4	GEN-005	MAT-003	GEN-003	MAT-003	GEN-007	MAT-001
5	GEN-010	MAT-011	GEN-010	MAT-002	GEN-002	MAT-011
6	GEN-017	MAT-002	GEN-002	MAT-017	GEN-004	MAT-002
7	GEN-004	MAT-012	GEN-004	MAT-004	GEN-002	GEN-002
8	GEN-011	MAT-011	GEN-020	MAT-017	GEN-020	GEN-020
9	GEN-010	MAT-017	GEN-002	MAT-002	GEN-010	MAT-012
10	GEN-002	MAT-002	GEN-024	MAT-012	GEN-004	MAT-002

NOTE: GEN = generalist, MAT = math content specialist.

Measures of Effective Teaching Sample

To determine rater assignments in the MET sample, a different design was used. First, we randomly created rater pairs (one content area specialist and one generalist). Then, we randomly assigned a block of teachers to each rater pair. Rater pairs scored both videos for each teacher. Blocks consisted of 10 teachers (math) and 8 teachers (ELA). There was one smaller block in each subject (N=8 math, N=7 ELA). Because of constraints on rater availability, in one math block, it was not possible to have a rater pair with both a content specialist and a generalist, so a random pair of two generalists was used to score videos. This resulted in a design that was fully crossed within blocks. Table 2.4 shows the assignment of raters using a scenario with 10 teachers and 20 math lessons and two teachers in each block. The design is similar for ELA lessons.

Table 2.4. Illustrative Rater Assignment Table for Math Teachers: Measures of Effective Teaching Sample

Teacher	Lesson 1		Lesson 2	
1	GEN-001	MAT-001	GEN-001	MAT-001
2	GEN-001	MAT-001	GEN-001	MAT-001
3	GEN-002	MAT-002	GEN-002	MAT-002
4	GEN-002	MAT-002	GEN-002	MAT-002
5	GEN-003	MAT-003	GEN-003	MAT-003
6	GEN-003	MAT-003	GEN-003	MAT-003
7	GEN-004	MAT-004	GEN-004	MAT-004
8	GEN-004	MAT-004	GEN-004	MAT-004
9	GEN-005	MAT-005	GEN-005	MAT-005
10	GEN-005	MAT-005	GEN-005	MAT-005

NOTE: GEN = generalist, MAT = math content specialist.

3. Methods

In this chapter, we provide details on the study methods used to address the two primary research questions. First, we describe our analytic methods for assessing score consistency and potential score bias. These analyses are critical to investigating the extent to which rater content expertise may influence the scoring process (research question 1), because they provide evidence about whether there are disagreements among raters, and whether these disagreements are systematic and related to rater background.

Second, we describe our methods for investigating score reliability. Establishing evidence of score reliability provides information about the extent to which scores can be used to characterize a teacher's instructional practice, broadly speaking. Reliability is increased to the extent that variation in observation scores reflects true variation in instructional practices, and diminished to the extent that variation in observation scores reflects variance from undesirable sources (also called error variance), including the sampling of lessons or the assignment of observers (Bell et al., 2012). If scores based on a small sample of lessons do not adequately characterize a teacher's instruction in other lessons, or if scores reflect rater error or differences in rater severity or leniency, it is difficult to provide useful feedback to teachers and difficult to understand how instructional practices relate to student performance (Haertel, 2013; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Third, we describe our methods for investigating the extent to which observation scores (overall and on each of the four rubric components) are valid predictors of teacher value-added scores in ELA and math (research question 2), or the extent to which TNTP Core scores extrapolate to other valued measures of teacher practice and quality (Bell et al., 2012). We describe the VAMs that were used to quantify teacher contributions to student achievement. Slightly different models were used across TNTP samples (sample 1 and sample 2) and the MET sample, and these models are detailed below. We then describe the meta-analytic framework we used to synthesize these relationships across districts and study samples.

Scoring Consistency and Potential Score Bias

We assess observers' agreement with one another using two different indices—*exact agreement* and *agreement-within-one*—and we also benchmark observer agreement with TNTP's 75/50 threshold. *Exact agreement* refers to the percentage of instances in which two raters agree completely in their scores for a particular lesson:

$$P = \frac{agreements}{agreements + disagreements} \times 100 \quad (1)$$

Agreement-within-one refers to the percentage of instances in which two raters assigned scores that were within one point of each other (e.g., if one rater assigned a score of 3 and another rater assigned a score of 4, this would constitute agreement-within-one):

$$P = \frac{(agreement\ within\ one)}{(agreement\ within\ one) + disagreements} \times 100 \quad (2)$$

As discussed in Chapter Two, TNTP's raters have to pass a calibration training by watching and rating five videos and having their scores meet the 75/50 rule. Across five videos, raters have to have a 75 percent agreement-within-one correspondence and a 50 percent exact agreement with an anchoring rating. We report a similar metric to TNTP's 75/50 threshold, which is the share of rated lessons that have 75 percent (three of four domains) with an agreement-within-one and 50 percent exact agreement, averaged across all lessons in our sample.

These indices provide information about the extent to which two raters agree with each other and, thus, the extent to which scores are expected to be consistently assigned regardless of which observer provided scores. The final rating, while unique to TNTP's training and rubric, provides information about the extent to which raters continue to meet the 75/50 rule after passing the initial calibration. Inconsistency could arise unsystematically—for example, it may be that raters seldom agree, but that, for some occasions, one rater provides a higher score than another rater, and on other occasions, the opposite is observed.

However, inconsistency could also arise systematically—it may be that one rater is systematically more severe than another rater and consistently provides lower ratings. This kind of inconsistency may represent a form of score bias. For example, if content specialists are systematically more lenient in their ratings than content generalists, this could create positive bias in scores for teachers who are always observed by content specialists. In this study, a particular source of bias that is of interest relates to rater background: Do observers with content area expertise provide systematically different ratings than observers with generalist backgrounds?

We investigate this by estimating the standardized mean difference between the scores assigned by generalist raters and the scores assigned by content specific raters:

$$d = \frac{\bar{X}_{SUB} - \bar{X}_{GEN}}{\sqrt{\frac{(n_{SUB} - 1)S_{SUB}^2 + (n_{GEN} - 1)S_{GEN}^2}{n_{SUB} + n_{GEN} - 2}}} \quad (3)$$

where $\bar{X}_{SUB}$ is the sample mean for the content specialist raters, $\bar{X}_{GEN}$ is the sample mean for the generalist raters, S_{SUB}^2 and S_{GEN}^2 are the sample variances for the content specialist and generalist raters, and n_{SUB} and n_{GEN} are the number of ratings provided by the content specialist and generalist raters. d is often referred to as Cohen's *d* (Cohen, 1977). As a rule of thumb, Cohen proposed that values of d less than 0.2 be considered small, and values greater than 0.8 be considered large.

However, it is not possible to fully characterize the bias in scoring based on the design of this study, because ultimately we cannot describe the accuracy of any one rater's score. For example, it may be that content specialists are more accurate than generalists, or vice versa. It may be that neither is accurate and that there are biases in both directions. To understand the nature of these systematic differences better and more fully characterize the nature of the bias, it would be necessary to compare both specialist and generalist ratings to "true" ratings provided by a master rater (which were not available in this study).

Reliability

We use a generalizability theory (G-theory) (Cronbach et al., 1972; Brennan, 2001) approach to appraise score reliability. G-theory provides a generalized framework for investigating reliability that allows for multiple sources of error to be quantified. Specifically, we employed an $(\boldsymbol{l}:\boldsymbol{t}) \times \boldsymbol{r}$ design to partition and quantify sources of variance in an observed TNTP Core score X_{tlr}, from one teacher (t) on a unique lesson (l) by one rater (r):

$$X_{tlr} = \mu + \mu_t + \mu_{l(t)} + \mu_r + \mu_{tr} + e_{tlr} \quad (4)$$

where μ is a grand mean, μ_t is a random effect for the teacher, $\mu_{l(t)}$ is a random effect for the lesson nested within teacher, μ_r is a random rater effect, μ_{tr} is a random rater-by-teacher interaction effect, and e_{tlr} is a residual error term. Each of these five random effects corresponds to a source of observation score variance, summarized in Table 3.1.

Table 3.1. Variance Components and Descriptors

Random Effect	Variance Component	Descriptor
μ_t	σ_T^2	Teacher variance.
$\mu_{l(t)}$	$\sigma_{l(t)}^2$	Variance due to lessons, confounded with teacher score dependence upon lessons.
μ_r	σ_r^2	Variance due to observer. Some observers are more stringent than others.
μ_{tr}	σ_{tr}^2	Variance due to teacher-rater interaction. Some teachers are scored higher by certain observers than others.
e_{tlr}	σ_{res}^2	Error variance confounded with teacher score dependence on observers and lessons.

σ_T^2 is the variance in observed teacher scores that reflects substantively meaningful differences in teacher instructional practice. This is the variance of interest. The other sources of variance constitute potential sources of measurement error. The score decomposition in Equation (4) allows us to quantify and appraise error variance from three sources: the sampling of lessons, the sampling of raters, and the rater-by-teacher interaction. The residual term captures error variance not accounted for in the model. In this model, it is not possible to separate out the lesson score variance from lesson score dependence on teacher—these two sources of variance are confounded.

As was described in Chapter Two, the study involved two separate samples of teachers—the TNTP sample and the MET sample. In the MET sample, the rater assignment was fully crossed within blocks. The full crossing allows for estimation of the variance sources in Table 3.1. In the TNTP sample, there are some teachers whose individual lessons were rated by different rater pairs, with no rater overlap between the lessons. In such a design, variance attributable to lessons is completely confounded with variance attributable to raters. However, there was a subset of teachers where at least one rater provided more than one lesson score. For these teachers, it is possible to disentangle rater and lesson effects. This subset of the TNTP sample was treated as an additional block, and the subdividing method (Chiu & Wolfe, 2002) was used to estimate variance components by pooling across blocks.

After partitioning and quantifying sources of variance, we used the results of the generalizability studies to estimate score reliabilities for different scenarios, based on the allocation of raters (n_r) and lessons (n_L):

$$\rho_T^2 = \frac{\sigma_T^2}{\sigma_T^2 + \frac{\sigma_{tr}^2}{n_r} + \frac{\sigma_{l(t)}^2}{n_L} + \frac{\sigma_{res}^2}{n_r n_L}} \quad (5)$$

We considered four different scoring designs: (A) one lesson scored by one rater, (B) four lessons scored by two raters, (C) two lessons scored by one rater, and (D) four lessons scored by one rater. These four designs represent common examples of how school systems may use observational rubrics to evaluate teachers. The results will provide evidence on the reliability of TNTP Core scores as schools would likely use them in an evaluation system.

Value-Added Models

Teacher VAMs aim to estimate the effect of teachers on their students' achievement by removing other factors related to student learning that are outside of teachers' control. Conceptually, these models generate a predicted score for students based on their prior achievement, their demographics, and the prior achievement and demographics of their peers and compare that with how students actually scored in a given year. Teachers with students who scored higher than predicted, on average, will have a positive value-added, and teachers with students who scored lower than predicted, on average, will have a negative value-added. A value-added of zero typically represents average growth for a given sample (e.g., school district or state).

In this section, we detail the VAMs that were used in our analyses. As we described in Chapter Two, we had to restrict the TNTP sample to estimate VAMs. TNTP sample 1 includes value-added scores for teachers across eight sites from a single state. These value-added scores were provided by the state department of education. TNTP sample 2 is drawn from a large charter network operating on the East coast with hundreds of teachers; the network was willing to share the necessary data for us to calculate VAMs. For this sample, we used all of the network's teachers to generate math and ELA value-added scores. We then restricted the sample to the 17 math teachers and 15 ELA teachers who had their instruction videotaped. In the MET sample, we used the value-added scores that were generated as part of the original MET study. Below, we describe the VAMs for each of the three samples.

Value-Added Model for TNTP Sample 1

The VAM estimates from TNTP sample 1 were provided to use by the state department of education. The state uses a residual-based approach to estimating value-added scores. First, students' current achievement is regressed on a vector of prior achievement and student demographics. Second, these residuals are averaged to the teacher level. The mean residual for a teacher represents the average difference between students' actual and predicted achievement for that teacher. These average residuals for teachers are considered their value-added scores, or their estimated effect on students' achievement, excluding factors outside teachers' control.

Value-Added Model for TNTP Sample 2

The vast majority of students in TNTP sample 2 were linked to multiple math and ELA teachers in the same year. Conceptually, the fact that one student has multiple teachers in the same year complicates the analysis. When the student has a single teacher, traditional value-added methods allocate all of the student's test scores to the teacher of record. However, it is less clear how the allocation should happen when the student has multiple teachers.

As a simple example, suppose that Teacher A and Teacher B co-teach a class consisting of nine students and that each teacher also has sole responsibility for teaching one additional student. Given this setup, we can consider a number of approaches to estimate the teacher's value-added score (VA). The first, termed the "Partial Credit Method" (see Hock & Isenberg, 2017), considers the overall effect of Teacher A and Teacher B on the students they co-taught to be equal to half of Teacher A's VA plus half of Teacher B's VA. Under this setup, while the scores of the students they jointly teach affect the sum of Teacher A's and Teacher B's VAs, those students' test scores do not affect the difference between the two teachers' VAs. To estimate the difference between Teacher A's and Teacher B's VAs, this approach relies solely on the two students who were not co-taught. In this approach, the difference between Teacher A's VA and Teacher B's VA is just the difference between the residualized score of the students that Teacher A taught alone and those whom Teacher B taught alone. While this approach is conceptually appealing, the VA estimates tend to be quite noisy.

Another approach, termed the "Full Roster Method" in Hock and Isenberg (2017), replicates the student observations so that they can be paired with every teacher who taught them; this means that the nine students in our hypothetical example above who were co-taught by the two teachers would each appear in the data twice, once matched to Teacher A and once matched to Teacher B. After this, the Full Roster Method uses a traditional technique for estimating VA, which is possible because each observation in the data is uniquely determined by a student-teacher pair. The technique accounts for the fact that the nine students were co-taught by weighting those students by half as much as the two students who were not co-taught in the resulting regressions. However, although this method uses weights to account for co-teaching, it implicitly treats the pairs (Teacher A and Teacher B) as independent. This means that Teacher B's VA will be biased in a way that depends on how effective Teacher A is; if Teacher A is a high-VA teacher, Teacher B's VA will be biased upward (i.e., credited for Teacher A's effectiveness), and if Teacher A is a low-VA teacher, Teacher B's VA will be biased downward (i.e., blamed for Teacher A's ineffectiveness).

The third approach, which we term the "Ridge Method," starts by approaching the problem in the same way as the Partial Credit Method. That is, we assume that the overall effect of Teacher A and Teacher B on the students they co-taught is equal to half of Teacher A's VA plus half of Teacher B's VA. However, instead of running a traditional ordinary least square regression using this specification, the Ridge Method minimizes the sum of squared errors plus the sum of the squared teacher VAs. This corresponds to a ridge regression, although one that does not include penalty terms of the non-value-added covariates (see Hastie, Tibshirani, & Friedman [2017] for more details about ridge regressions).

The Ridge Method can be motivated by using the Bayesian framework, since the resulting VA estimates correspond to a Bayesian best estimate for each teacher's value-added score. In addition, there are some additional aspects of the estimator that makes us prefer the Ridge Method over the other two methods discussed above. Like the Full Roster Method, the Ridge Method prefers to apportion the credit equally to Teacher A and B for the test score gains or losses realized by the co-taught students, which leads to much more-precise VA estimates than the Partial Credit Method.[7] However, unlike the Full Roster Method, the Ridge Method still accounts for who the other teacher is in the co-taught classroom, which reduces the bias of the VA estimates. For example, suppose that Teacher A has been teaching for many years, so we feel confident in knowing that he or she has an above-average VA. The Ridge Method incorporates this information into the estimates of Teacher B's VA, acknowledging the fact that the nine students who were co-taught would probably do well even if Teacher B is an average teacher, because Teacher A is an above-average teacher. The Full Roster Method, in contrast, ignores this information and produces teacher VAs for Teacher B that are (roughly) the average value-added scores for both teachers, regardless of how much information is available about Teacher A's quality.

Based on the favorable properties of the Ridge Method, we employ it to estimate VAs separately for each subject and allow for each teacher to have different VAs in each year. As covariates, we include a cubic function of the student's prior test scores in the same subject (and off subject), as well as variables that indicate whether the student is male or female, on free or reduced-price lunch, classified as having limited English proficiency, and classified as being special education. In our ridge regression, we use a penalization term of 25; we chose this because, in the Bayesian framework, 25 corresponds to a prior distribution of the VAs with a mean of zero and standard deviation (SD) of 0.2, which roughly corresponds to the VA distribution found in other studies (e.g., Chetty, Friedman, & Rockoff, 2014).

Value-Added Model for Measures of Effective Teaching Sample

For the MET sample, we used the 2010 math and ELA value-added scores already generated by the study team (Kane et al., 2013). The VAM follows the same logic described above, and the details are available in this report. Because our sample was restricted to 4th- and 5th-grade teachers, teachers in our sample were responsible for only a single class (or *section*, as referred to in the MET study). Furthermore, while the majority of the teachers we sampled from the MET

[7] The Ride Method prefers to give Teacher A and Teacher B equal credit or blame for the students they co-taught, because apportioning the credit equally minimizes the sum of the squared VA terms.

study had both math and ELA value-added scores, we use teachers' value-added scores for the subject in which they were assigned for the rater portion of the study (e.g., if teachers' were assigned to be watched by ELA raters, we used only their ELA value-added scores).

Summarizing Effects with Meta-Analysis

Once we had value-added scores for all available teachers, we estimated first-order Pearson correlations with math and ELA value-added scores, TNTP Core domain scores, and overall TNTP Core scores. We estimated these correlations and accompanying standard errors (SEs) using Equation (6):

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}; \quad se_{r_{xy}} = \sqrt{\frac{1-r_{xy}^2}{n-2}}, \qquad (6)$$

where r_{xy} is the Pearson correlation coefficient between a TNTP Core domain or overall rating (x) and math or ELA value-added (y), σ_{xy} is the covariance between TNTP Core domain or overall rating (x) and math or ELA value-added (y), and σ_x and σ_y are the respective standard deviations. We ran these correlations separately for TNTP sample 1, TNTP sample 2, and the MET sample.

In prior research, including the MET Study (Kane & Staiger, 2013), researchers have disattenuated the correlations between value-added and other measures of teacher practice. The logic is that both measures have error, and this error pushes (or attenuates) the correlation toward zero. Correcting the correlations for measurement error, or disattenuating them, potentially provides a better estimate of the true correlation between the measures of teacher quality. However, this often relies on untenable assumptions. For example, in the MET study, the disattenuation relies on the assumption that the disattenuated correlations represent the relationship between a value-added score and an observational rubric score that was measured by an infinite number of observations by an infinite number of observers. In short, these disattenuated correlations do not provide policy-relevant estimates when most teachers are observed three to five times by a single reviewer each time. For this reason, we report only unadjusted correlations. With the site-specific Pearson correlations in hand, we combine the estimates into an overall correlation, standard error, and *p*-value using a fixed-effect meta-analysis (see Borenstein et al., 2010).

4. Consistency, Bias, and Reliability

In this chapter, we describe evidence of the extent to which scoring procedures are consistent, the extent to which scores are bias-free, and the extent to which scores reliably characterize a teacher's instructional practice. We also present descriptive statistics for TNTP Core scores summarizing score distributions.

Descriptive Statistics

Means and standard deviations, as well as correlations among the four domains, are summarized in Table 4.1 The descriptive statistics presented in Table 4.1 represent the distribution of teacher scores, averaged across lessons and raters. The means were relatively close to the midpoint of the scale, though the mean for Culture of Learning was high in both math and ELA. The teacher scores did not vary greatly on any of the dimensions, and in fact there was some range restriction in scores—no teachers had average scores at the highest level (skillful) on any domain. In both subjects, domain scores were moderately to strongly correlated with each other. In particular, Academic Ownership and Demonstration of Learning scores were very highly correlated, suggesting that teachers who encourage students to take responsibility for their learning also tend to have all students demonstrate that they are learning. Across both subjects, the mean score within each domain was at or just below the midpoint of the scale (e.g., a score of 3), which is potentially due to the low-stakes nature of our research design.

Table 4.1. Descriptive Statistics and Correlations Among Dimension Scores

Math						
				Correlation Among Domains		
Domain	**Mean**	**SD**	**Range**	**2.**	**3.**	**4.**
1. Culture of Learning	3.26	0.64	[1.50,4.50]	0.61	0.73	0.74
2. Essential Content	2.69	0.74	[1.00,4.00]	1.00	0.67	0.70
3. Academic Ownership	2.46	0.68	[1.00,4.17]		1.00	0.89
4. Demonstration of Learning	2.62	0.65	[1.25,4.17]			1.00
ELA						
				Correlation Among Domains		
Domain	**Mean**	**SD**	**Range**	**2.**	**3.**	**4.**
1. Culture of Learning	3.19	0.54	[1.33,4.00]	0.55	0.48	0.64
2. Essential Content	2.68	0.64	[1.25,4.00]	1.00	0.73	0.74
3. Academic Ownership	2.52	0.62	[1.50,3.83]		1.00	0.81
4. Demonstration of Learning	2.55	0.57	[1.50,4.00]			1.00

NOTE: Sample includes TNTP Core Samples 1 and 2 plus the MET sample; Math N=67, ELA N=66.

Figure 4.1 and Figure 4.2 show the score frequencies for each domain in math and ELA, respectively, for teachers in TNTP Core samples 1 and 2 plus the MET sample. Scores were not evenly distributed and tended to be right-skewed. Very few scores of skillful were assigned to any lessons by any raters (less than 2 percent overall). Approximately 90 percent of all scores were between 2 (minimally effective) and 4 (proficient).

Figure 4.1. TNTP Core Dimension Frequencies, Math

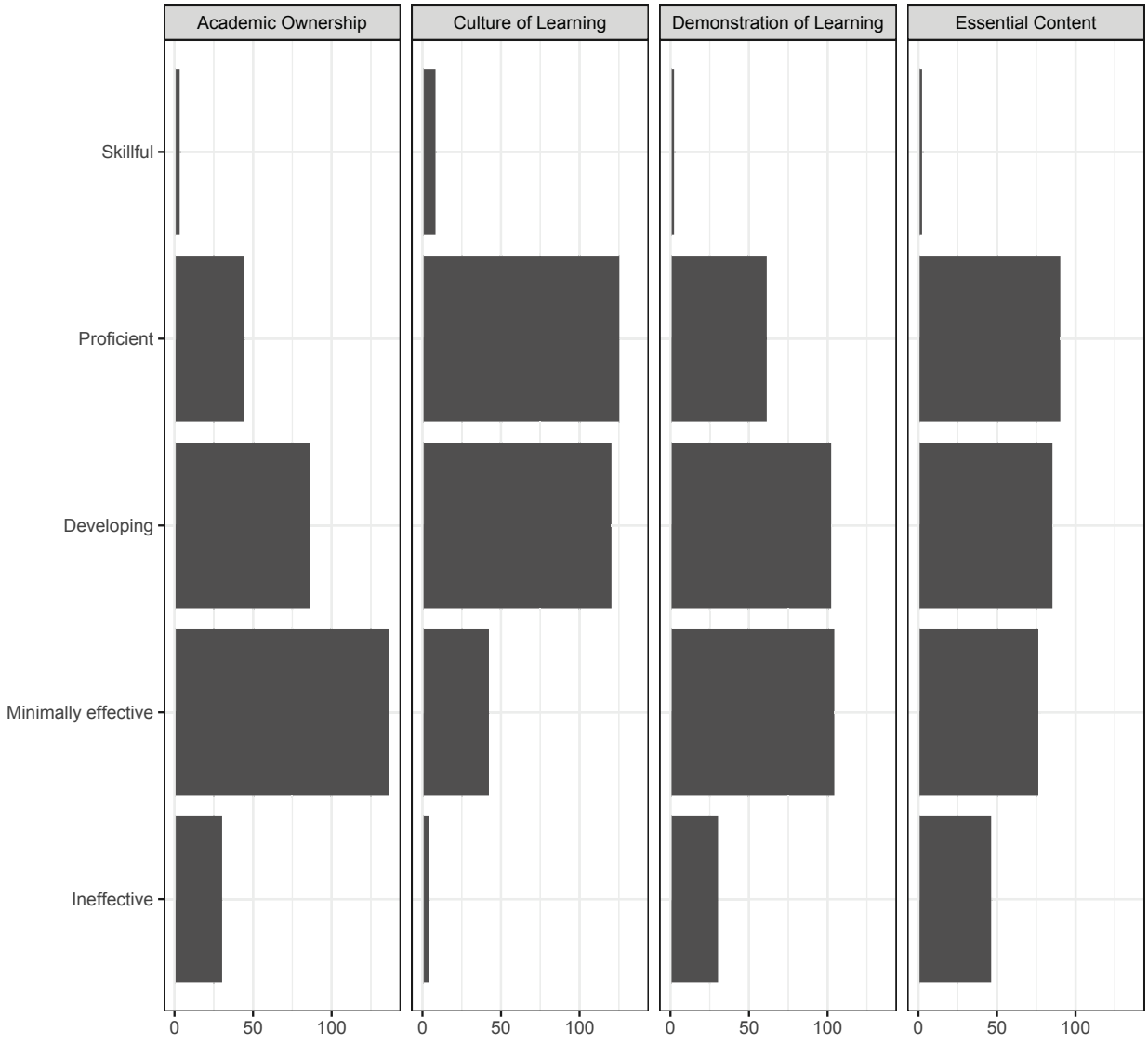
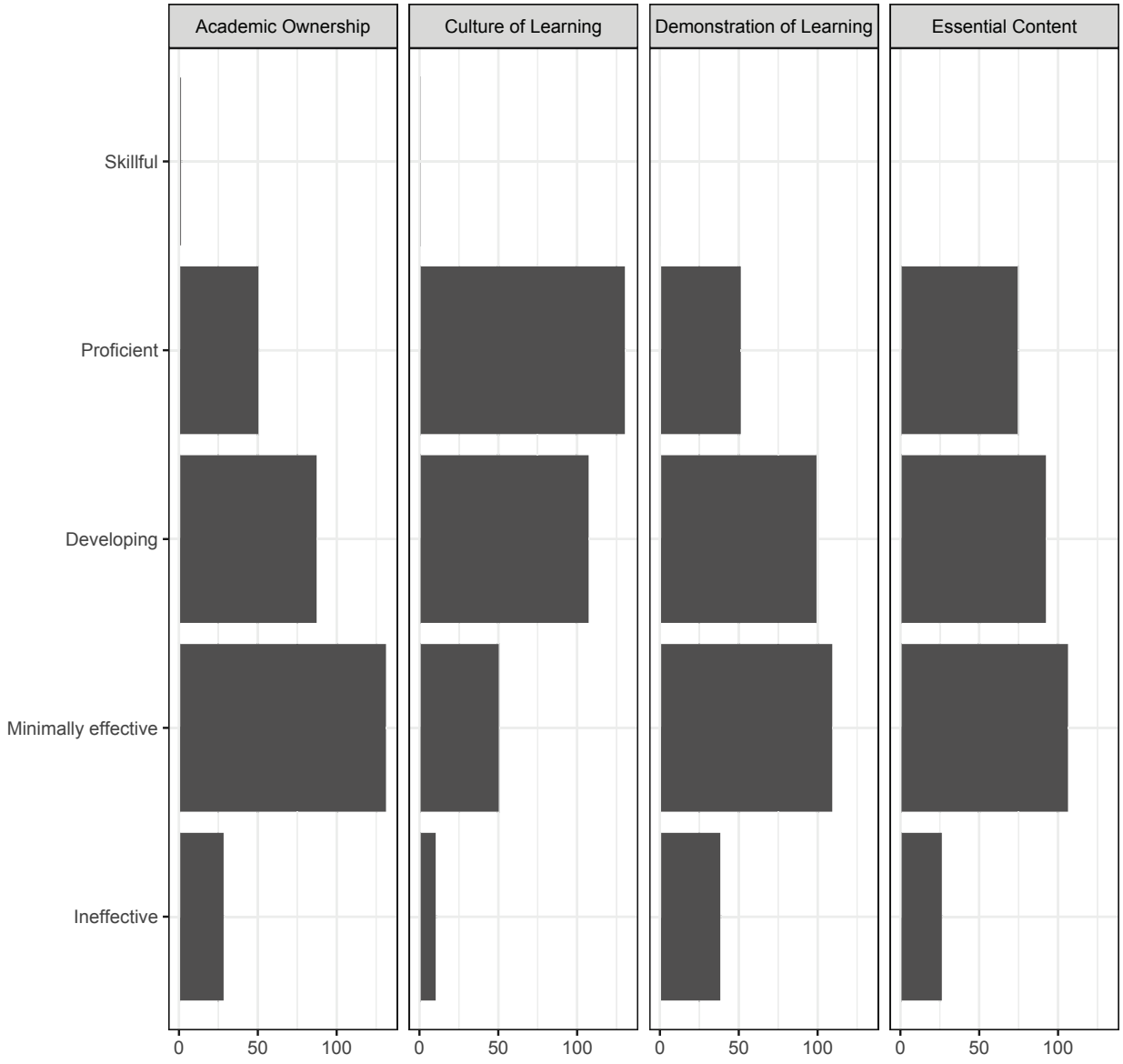


Figure 4.2. TNTP Core Dimension Frequencies, ELA



Scoring Consistency and Potential Score Bias

Rater Agreement

Exact rater agreement varied between 28 and 45 percent, depending on the domain, and agreement-within-one ranged from 76 percent to 94 percent (Table 4.2). Culture of Learning had the highest level of exact agreement for math and the lowest level of exact agreement for ELA. While there are not clear standards to benchmark these agreement rates (Bell et al., 2012), these agreement rates are not adjusted for chance agreement and reflect substantial rater disagreement in the scoring process. Using TNTP's benchmark of 75 percent agreement-within-one and 50 percent exact agreement, less than half of all scored lessons meet this criteria for acceptable agreement (Table 4.3). While the 75/50 metric measures disagreement between two raters (not

between a rater and a master rater) and is averaged across the sample, the results in Table 4.3 indicate that raters in the field drift away from their calibration training.

Table 4.2. Rater Agreement Statistics

	Math		ELA	
Domain	**% Exact**	**% One Off**	**% Exact**	**% One Off**
Culture of Learning	45.39	94.08	30.46	85.43
Essential Content	28.29	78.95	36.42	80.13
Academic Ownership	38.16	91.44	36.42	83.44
Demonstration of Learning	32.89	82.89	36.42	76.16

Table 4.3. Share of Lessons Meeting 75/50 Threshold

Subject	% of Lessons Meeting 75% Threshold	% of Lessons Meeting 50% Threshold	% of Lessons Meeting Both Criteria
Math	88.16	47.37	47.37
ELA	83.44	45.70	45.70

Differences by Rater Background

There was some evidence, particularly in math, that there may be bias in the scoring of lessons. As shown in Table 4.4, there are systematic differences in rater severity by rater type. For ELA, content specialists tended to rate lessons more leniently on all domains than content generalists, though the standardized differences between the ratings of the two types of raters are relatively small (i.e., less than 0.2), particularly in Essential Content. For math, content specialists tended to rate lessons more severely on all domains, and the standardized differences are moderate to large. On Essential Content, math content specialists rated teachers over a half of a standard deviation lower than their generalist peers. These differences are important to keep in mind for the use of TNTP Core to evaluate teacher performance, as teachers consistently observed by math specialists may receive lower ratings than their peers observed by generalist raters.

Table 4.4. Comparing TNTP Core Scores by Rater Type

	Math		
	Specialist	Generalist	Standardized
Domain	Mean (SD)	Mean (SD)	Difference
Culture of Learning	3.26 (0.85)	3.37 (0.78)	–0.13
Essential Content	2.46 (1.11)	2.99 (0.97)	–0.51
Academic Ownership	2.34 (0.94)	2.78 (0.87)	–0.49
Demonstration of Learning	2.54 (0.99)	2.84 (0.89)	–0.32
	ELA		
	Specialist	Generalist	Standardized
Domain	Mean (SD)	Mean (SD)	Difference
Culture of Learning	3.26 (0.81)	3.13 (0.88)	0.15
Essential Content	2.74 (0.90)	2.69 (0.98)	0.05
Academic Ownership	2.62 (0.88)	2.48 (0.92)	0.16
Demonstration of Learning	2.66 (0.92)	2.45 (0.94)	0.23

NOTE: Sample includes TNTP Core samples 1 and 2 and the MET sample; N Math = 134; N ELA = 151. Difference is the standardized mean difference.

Reliability

Variance decompositions for math classrooms are summarized in Table 4.5. Most of the variation in TNTP Core scores is not attributable to teachers, and there is considerable variation across lessons, raters, and other unexplained sources. Teachers contributed between 6 and 31 percent of the variance in scores, depending on the domain. This result suggests that for some domains—in particular, the Academic Ownership domain—there is not much variability in practices, on average, across teachers. There were sizable differences in scores across lessons—between 9 and 12 percent of the variance, depending on domain—suggesting that scores are likely to vary depending on which lessons are sampled and that generalizing from a single observation will likely involve uncertainty. Raters accounted for between 19 and 31 percent of the variance, depending on domain.

Table 4.5. Variance Decomposition and Implied Reliability, Math

	Percentage of Variance						Reliability			
Domain	σ_t^2	$\sigma_{l(t)}^2$	σ_r^2	σ_{tr}^2	σ_{res}^2	SE	A	B	C	D
Culture of Learning	30.64	8.81	19.07	15.52	25.96	0.54	0.38	0.48	0.56	0.70
Essential Content	10.34	10.63	28.1	7.8	43.13	0.81	0.14	0.23	0.33	0.46
Academic Ownership	5.77	11.63	26.97	20.22	35.41	0.64	0.08	0.12	0.15	0.25
Demonstration of Learning	10.2	9.26	31.29	10.32	38.94	0.67	0.15	0.23	0.31	0.45

NOTE: Standard error based on one rating by one observer. A: 1 rater, 1 lesson. B: 1 rater, 2 lessons. C: 1 rater, 4 lessons. D: 2 raters, 4 lessons

The implied reliability of scores for each of the four scoring designs (Table 4.5) we investigated highlights the fact that in order to support generalizations from observation scores to a teacher's instructional practice more broadly, scores would be needed from multiple raters over multiple lessons. With only one rater on one occasion, these coefficients are low, ranging between 0.08 and 0.38. This is more easily seen when the reliability estimates are translated into standard errors of measurement. For example, the standard error of measurement based on one rater and one lesson is approximately 0.81 for the Essential Content domain (or nearly a full point on TNTP Core). This means that there is sufficient uncertainty in the score so that a teacher who is rated "developing" could actually be either "proficient" or "minimally effective."

With two raters observing four different lessons, the estimated reliability coefficients increase quite a bit (between 85 percent and 230 percent), although even with this many observations, the generalizability coefficients are relatively low, in an absolute sense.

Similar results were observed for ELA classrooms (Table 4.6). Teachers contributed between 7 and 22 percent of the variance in scores, depending on the domain. For the Demonstration of Learning domain, there is not much variability in practices, on average, across teachers. There were differences in scores across lessons—between 5 and 14 percent of the variance, depending on domain. Raters accounted for the largest portion of observation score variance—between 34 and 47 percent of the variance, depending on domain—meaning that teacher scores depend significantly on who performed the observation.

Table 4.6. Variance Decomposition and Implied Reliability, ELA

	Percentage of Variance						Reliability			
Domain	σ_t^2	$\sigma_{l(t)}^2$	σ_r^2	σ_{tr}^2	σ_{res}^2	SE	A	B	C	D
Culture of Learning	18.67	5.12	46.80	3.91	25.49	0.55	0.35	0.49	0.62	0.74
Essential Content	21.77	8.32	33.51	13.52	22.86	0.67	0.33	0.43	0.51	0.65
Academic Ownership	16.45	9.81	45.33	11.32	17.09	0.58	0.30	0.40	0.48	0.62
Demonstration of Learning	6.52	14.17	46.93	13.59	18.78	0.71	0.12	0.18	0.23	0.34

NOTE: Standard error based on one rating by one observer. A: 1 rater, 1 lesson. B: 1 rater, 2 lessons. C: 1 rater, 4 lessons. D: 2 raters, 4 lessons

Again, consistent with math results, the implied reliability of scores for each of the four scoring designs (Table 4.6) we investigated highlights the fact that in order to support generalizations from observation scores to a teacher's instructional practice more broadly, scores would be needed from multiple raters over multiple lessons. With only one rater on one occasion, these coefficients are low, ranging between 0.12 and 0.35. With two raters observing four different lessons, coefficients range from 0.34 to 0.74.

Overall, these results suggest that inferences about overall teacher instructional practice based on scores provided by a limited number of raters from a limited sample of lessons are tenuous at best, and using these scores to characterize instructional practice should be done cautiously.

5. Extrapolation and Predictive Validity

The second research question addresses the extent to which teachers' TNTP Core scores extrapolate to other related measures of teacher quality or practice (Bell et al., 2012). TNTP Core was designed to measure teaching practices that affect a variety of student outcomes, but especially students' learning. In our study, we measure the relationship between TNTP Core scores and student learning by correlating teachers' aggregate rubric scores (averaged across lessons and raters) with their value-added scores. As described above, teachers' value-added scores aim to capture teachers' independent effects on students' learning as measured by math and ELA assessments.

Subject-Specific Correlations

Math

In this section, we correlate teachers' overall scores and four domain TNTP Core scores with their math value-added scores, separately by site. We also pool the site-specific correlations using a fixed-effect meta-analysis. The results are reported in Table 5.1. In general, across teachers' overall scores and domain scores, there is not a strong pattern between TNTP Core and teachers' math value-added. Across the three sites, the Academic Ownership domain has the strongest positive relationship with value-added (meta-analytic $r_{xy} = 0.122$) but does not reach standard levels of statistical significance. While the overall and the domain-specific TNTP Core scores (except for Essential Content) have small positive correlations with math value-added, the relationships and inference do not rule out true correlations between –0.2 and 0.2. In short, in our study sample there is no clear statistically significant relationship between teachers' overall and domain TNTP Core scores and students' math learning as measured by value-added scores. It should be noted that our small sample size limits our statistical power and ability to find statistically significant correlations among TNTP Core scores and teachers' value-added scores.

Table 5.1. Pearson Correlations of Math Value-Added and Teachers' Math TNTP Core Scores

	Math							
	TNTP Sample 2 (N=17)		MET Sample (N=48)		TNTP Sample 1 (N=5)		Fixed-Effect Meta-Analysis	
Domain	Pearson Correlation	SE	Pearson Correlation	SE	Pearson Correlation	SE	Pearson Correlation	SE
Overall	0.432+	0.233	–0.053	0.147	–0.033	0.577	0.080	0.122
Culture	0.368	0.240	–0.009	0.147	–0.269	0.556	0.077	0.123
Essential	0.450+	0.231	–0.186	0.145	0.099	0.574	–0.001	0.120
Academic	0.215	0.252	0.091	0.147	0.117	0.573	0.122	0.124
Demonstration	0.456+	0.230	–0.050	0.147	0.003	0.577	0.093	0.121

NOTE: + $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, and *** $p \leq 0.001$.

English Language Arts

Contrary to the relationship between the math TNTP Core scores and value-added, the ELA relationship suggests a small-to-moderate positive relationship between teachers' TNTP Core scores and ELA value-added (Table 5.2). Across the three sites, correlations between teachers' ELA value-added scores and the overall TNTP Core score or the domain-specific TNTP Core scores were consistently positive (with the one exception of the Academic Ownership domain for MET sample, where there was a small, negative correlation with value-added scores). When the correlations are pooled across sites, all but Academic Ownership has a correlation in the 0.17 to 0.22 range. These correlations are similar in magnitude to prior studies, including MET (Gill et al., 2017; Kane & Staiger, 2013). Further, we have reported raw correlations between teachers' TNTP Core scores and ELA value-added, unadjusted for attenuation bias. Although our study has a limited sample size, our results suggest that the overall TNTP Core score, and three of the four domains moderately predict student ELA learning as measured by teachers' ELA VAMs.

Table 5.2. Pearson Correlations of ELA Value-Added and Teachers' ELA TNTP Core Scores

	ELA							
	TNTP Sample 2 (N=15)		MET Sample (N=45)		TNTP Sample 1 (N=25)		Fixed-Effect Meta-Analysis	
Domain	Pearson Correlation	SE	Pearson Correlation	SE	Pearson Correlation	SE	Pearson Correlation	SE
Overall	0.367	0.258	0.125	0.151	0.131	0.207	0.171	0.110
Culture	0.049	0.277	0.288+	0.147	0.095	0.208	0.196+	0.110
Essential	0.479+	0.243	0.209	0.149	0.047	0.208	0.219*	0.109
Academic	0.345	0.260	-0.136	0.152	0.143	0.206	0.031	0.110
Demonstration	0.380	0.257	0.098	0.151	0.202	0.204	0.180^	0.110

NOTE: + $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, and *** $p \leq 0.001$; ^ p-value = 0.102.

Correlations Pooled Across Subjects

Given that the goal of TNTP Core is to measure generalizable teaching practices, we also provide correlations pooled across subjects (math and ELA). We follow a similar approach to the one used for subject-specific pooled correlations: We first pool our math and ELA correlations within a site, and we then pool the site-specific correlations across sites using a fixed-effect meta-analysis. In Table 5.3, we report the pooled correlations (across subjects and sites). As predicted by the subject-specific results, the pooled correlations are in between the math and ELA meta-analytic correlations. With the exception of Academic Content, the overall and three domains have pooled (across subjects and sites) correlations between 0.12 and 0.14. These correlations also do not meet conventional levels of statistical significance, but given our small sample size, they suggest that TNTP Core scores have small positive correlations with math and ELA value-added scores, when pooled across subjects.

Table 5.3. Meta-Analysis Correlations Pooled Across Math and ELA

Pooled Across ELA and Math			
	Fixed-Effect Meta-Analysis		
Domain	Pearson Correlation	SE	*p*-value
Overall	0.130	0.082	0.111
Culture	0.143	0.082	0.081
Essential	0.120	0.080	0.136
Academic	0.072	0.083	0.383
Demonstration	0.141	0.081	0.084

6. Conclusion and Recommendations

Conclusion

Taken together, these results have several implications for schools and school districts intending to use TNTP Core to provide teachers with feedback on their performance. We emphasize three key points here.

Relationships between teachers' TNTP Core scores and student achievement gains are modest and vary by subject area.

We use teachers' math and ELA value-added scores as a measure of instructional quality. Value-added scores capture teachers' contributions to students' math and ELA achievement, excluding factors outside teachers' control. Our results suggest that TNTP Core scores correlate with teacher value-added scores, though the strength of these relationships differs by subject area and by TNTP Core domain. In math, we find correlations between 0 and 0.12, and in ELA, we find correlations between 0.03 and 0.21. A correlation of 0.2 suggests that TNTP Core scores explain just 4 percent of teachers' value-added score ranking.

Observers often disagree in their ratings of instructional practice based on TNTP Core, and these disagreements may be related to their content expertise.

Across all four domains and subjects, on average, raters struggled to agree on their judgments about the quality of a lesson. Overall, agreement rates were no better than what would be expected by chance. Raters also varied in the severity of lesson ratings in math, and this was in part based on content knowledge. Math specialist raters in particular were more likely to rate a lesson lower than their generalist peers.

There is a considerable amount of uncertainty about the extent to which TNTP Core scores represent teachers' overall instructional practices.

Disagreements among raters introduce uncertainty into TNTP Core scores and weaken claims about teacher practices that can be made on the basis of those scores. However, raters are not the only source of uncertainty in TNTP Core scores. There is also uncertainty that is introduced based on which specific lessons are observed and rated. To make general claims about a teacher's instructional quality from TNTP Core scores, one would need many observations of a given teacher conducted by many raters throughout the year. Our results suggest that even if teachers were observed four times per year, each time by two raters, there would still be considerable uncertainty about the quality of teachers' instructional practice. This uncertainty has at least two consequences. First, it attenuates relationships between TNTP Core scores and other indicators of teacher quality, making it difficult to assess the extent to which the aspects of instruction measured by TNTP Core are related to student outcomes. Second, it diminishes the usefulness of feedback that can be given based on TNTP Core scores. If TNTP Core scores do not accurately represent the practices of teachers overall, then their utility in diagnosing and developing instructional practices is compromised.

Recommendations

We conclude with a number of recommendations for schools, school districts, and charter networks intending to use TNTP Core, focusing on actions that may improve the quality of feedback that can be provided based on TNTP Core scores.

Set High Standards for Rater Certification

Currently, TNTP uses a 75/50 threshold for rater certification. In order to be considered qualified to use TNTP Core to rate instruction, across five lessons trainees must reach 75 percent agreement-within-one and 50 percent exact agreement with a master rater. We suggest raising the criteria scores for rater certification as one potential method to increase rater agreement. While meeting certification requirements is no guarantee that, in practice, raters will rate consistently or accurately, it is possible that higher certification standards can promote increased rater agreement. At a minimum, criterion scores should account for the possibility of chance agreement. For this reason, adopters may even want to consider using 90 percent agreement-within-one and 70 percent exact agreement as criteria scores for rater certification, in line with other certification processes (Kane & Staiger, 2012; Bell et al., 2012).

Build Systems to Monitor the Score Quality Under Operational Conditions

Many factors influence the quality of observation scores, including the context in which scores are collected, the extent to which teachers vary in their instructional practices, and the extent to which raters apply scoring rules consistently and accurately. While initial certification and training are an important component in ensuring that scoring rules are applied with fidelity, research has shown that raters tend to change their approach to scoring over time—a phenomenon known as rater drift. Rater drift begins to occur almost immediately after training and continues over time (Casabiana, Lockwood, & McCaffrey, 2014). It may be helpful for schools and school districts to embed more frequent post-certification calibration and validation exercises during a rating period (Bergin et al., 2017). Calibration could involve having all participating raters score the same (master-rated) lesson video and discuss score inconsistencies. Validation could involve monitoring raters' scores by having a master rater score a subset of the lesson videos. It may be helpful to conduct these activities as often as weekly (Bell et al., 2012).

High-Quality Evaluation and Feedback Requires Many Observers with Different Backgrounds to Rate Many Lessons

Consistent with other research (Kane & Staiger, 2012), we found that assertions about a teacher's overall instructional practices based on TNTP Core scores would be valid only if they were based on many observations of a given teacher conducted by many raters throughout the year. There is likely to be considerable uncertainty about a teacher's overall instructional practices even when assertions about instructional practice are based on four observations conducted by two raters in a given school year. Our findings suggest that scores also depend on the specific rater, because content expertise can influence the severity or leniency of ratings. This suggests not only that many raters be used, but also that it is important to ensure that teachers are observed by both content experts and content generalists. It may also be helpful to administer a content knowledge assessment (in this study, we used the Content Knowledge for Teaching

[CKT]) to all prospective raters, in order to gain some insight into content expertise prior to engaging in classroom observations.

Consider Collecting Additional Evidence of Validity

Reliability and validity are not static properties of observation protocols; they are better thought of as processes—they depend, for example, on who is observed and the conditions under which scores are collected. Scores collected from researchers or external observers may have different properties than scores collected from principals or peers. Additionally, validity is closely tied to how scores are intended to be used. Scores could potentially be used in a wide range of ways: to promote conversations in grade level or content area teams, to guide or inform professional development plans, or to guide or inform decisions during performance evaluations. Different kinds of evidence would be necessary to support these intended uses, and uses that are tied to consequences have lower tolerance for uncertainty. As schools, school districts, and charter networks outline specific uses for TNTP Core, they should consider collecting other sources of evidence that support claims about the quality of teacher practice—including evidence based on the content of the protocol (Does it adequately measure the intended dimensions?), evidence based on outcomes and consequences, and evidence based on response processes (how raters are interpreting and scoring the domains). If TNTP Core is used as one measure in a multiple-measure system, separate validity evidence should be collected supporting inferences based on multiple measures.

7. Limitations

The goal of the study is to produce rigorous evidence that TNTP Core scores are reliable and support valid inferences about instructional practice. Specifically, the study aims to measure the effect of raters and rater background on TNTP Core scores, as well as the predictive validity of these scores on students' achievement. The design of the study aimed to remove a number of confounding factors that may impede our ability to answer these questions. However, our study has a number of limitations that are worth noting.

First, evidence of validity and reliability is always context specific (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Our samples include students and teachers from traditional public schools and charter schools across a number of states and school years. However, our sample is not a random sample of all students and teachers in the United States. Additionally, while the inclusion of the MET Sample increased our study sample size and provided robust estimates of variance components for our analyses of score consistency, accuracy, and generalizability, there are a few issues with this sample that merit consideration. First, TNTP Core was designed to be sensitive to CCSS-aligned instructional practices, and the MET Sample was collected before CCSS implementation. To the extent that skillful practice (as measured by TNTP Core) is highly aligned to practices that are uniquely associated with CCSS implementation, it is possible that there is a misalignment between the kinds of practices captured by TNTP Core and the kinds of practices captured in the MET videos, and this can influence conclusions about score generalizability. It is important to keep these considerations in mind when applying the results of this study to other sites or contexts.

Second, the VAMs were not the same across all three samples. While the basic setup of each VAM was conceptually similar, each model had unique attributes. First, we were unable to generate our own VAM in TNTP sample 1. Second, we had to correct for the assignment of multiple teachers to each student in TNTP sample 2. Third, the MET study used yet another model that differed from the first two samples. Finally, each site also used a different math and ELA assessment. It is possible that the instructional sensitivity and other measurement properties differ across these sites. We try to account for these differences by first estimating correlations within each sample and combining sample-specific estimates using a fixed-effect meta-analysis. Furthermore, the sample for whom TNTP Core scores and value-added scores were available was small.

Third, our research design is focused on three aspects of validity: scoring consistency and bias, generalization of rater scores, and extrapolation of rater scores to other measures of teacher quality. Our research design does not cover all aspects of validity important to the design and implementation of TNTP Core. For example, we did not conduct a series of cognitive interviews to assess whether raters interpret the rubric as intended by TNTP. Given that TNTP Core is unique in its focus on student behavior, as opposed to teacher behavior, a cognitive interview exercise may yield useful information about the design of TNTP Core. Furthermore, we cannot speak to another unique element of TNTP Core: Each domain score is derived from a single

item, in contrast to other rubrics that often use multiple items per domain. Our research design is unable to assess whether the single-item aspect of TNTP Core influences how raters evaluate lessons.

References

Achieve, *Strong Standards: A Review of Changes to State Standards Since Common Core*, Washington, D.C., 2017. As of February 27, 2018: https://www.achieve.org/files/StrongStandards.pdf

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, 2014.

Ball, D. L., and Forzani, F. M., "Building a Common Core for Learning to Teach: And Connecting Professional Learning to Practice," *American Educator*, Vol. 35, No. 2, 2011, pp. 17–21.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., and Qi, Y. "An Argument Approach to Observation Protocol Validity," *Educational Assessment*, Vol. 17, No. 2–3, 2012, pp. 62–87.

Bergin, C., Wind, S. A., Grajeda, S., and Tsai, C.-L., "Teacher Evaluation: Are Principals' Classroom Observations Accurate at the Conclusion of Training?" *Studies in Educational Evaluation*, No. 55, 2017, pp. 19–26.

Borenstein, M., Hedges, L. V., Higgins, J., and Rothstein, H. R., "A Basic Introduction to Fixed-Effect and Random-Effects Models for Meta-Analysis," *Research Synthesis Methods*, Vol. 1, No. 2, 2010, pp. 97–111.

Brennan, R. L., *Generalizability Theory*, New York: Springer-Verlag, 2001.

Casabianca, J. M., Lockwood, J. R., and McCaffrey, D., "Trends in Classroom Observation Scores," *Educational and Psychological Measurement*, Vol. 75, No. 2, 2014, pp. 311–337.

Chetty, R., Friedman, J. N., and Rockoff, J. E., "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, Vol. 104, No. 9, 2014, pp. 2593–2632.

Chiu, C. W. T., and Wolfe, E. W., "A Method for Analyzing Sparse Data Matrices in the Generalizability Theory Framework," *Applied Psychological Measurement,* Vol. 26, No. 3, 2002, pp. 321–338.

Cohen, J., *Statistical Power Analysis for the Behavioral Sciences*, revised edition, Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc., 1977.

Cohen, J., and Goldhaber, D., "Building a More Complete Understanding of Teacher Evaluation Using Classroom Observations," *Educational Researcher*, Vol. 45, No. 6, 2016*,* pp. 378–387.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N., *The Dependability of Behavioral Measurements*, New York: Wiley, 1972.

Doherty, K. M., and Jacobs, S., "State of the States 2013 Connect the Dots: Using Evaluations of Teacher Effectiveness to Inform Policy and Practice," National Center on Teacher Quality, 2013.

Gill, B., Shoji, M., Coen, T., and Place, K., *The Content, Predictive Power, and Potential Bias in Five Widely Used Teacher Observation Instruments*," Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory, REL 2017–191, 2016.

Goe, L., Bell, C., and Little, O., *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis*, Washington, D.C.: National Comprehensive Center for Teacher Quality, 2008.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., and Schuermann, P., "Make Room Value Added: Principals' Human Capital Decisions and the Emergence of Teacher Observation Data," *Educational Researcher*, Vol. 44, No. 2, 2015, pp. 96–104.

Haertel, E., *Reliability and Validity of Inferences About Teachers Based on Student Test Scores (Angoff Lecture)*, Princeton, N.J.: Educational Testing Service, 2013.

Hastie, T., Tibshirani, R. and Friedman, J., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," *Springer Series in Statistics*, 2017.

Hock, H., and Isenberg, E., "Method for Accounting for Co-Teaching in Value-Added Models," *Statistics and Public Policy*, Vol. 4, No. 1, 2017, pp. 1–11.

Hsieh, W. Y., Hemmeter, M. L., McCollum, J. A., and Ostrosky, M. M., "Using Coaching to Increase Preschool Teachers' Use of Emergent Literacy Teaching Strategies," *Early Childhood Research Quarterly*, Vol. 24, No. 3, 2009, pp. 229–247.

Kane, M., "*Validation*," in R. Brennan, ed., *Educational Measurement*, 4th ed., Westport, Conn.: American Council on Education and Praeger, 2006, pp. 17–64.

Kane, T. J., McCaffrey, D. F., Miller, T., and Staiger, D. O., "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment," 2013. As of August 14, 2018: http://k12education.gatesfoundation.org/download/?Num=2676&filename=MET_Validating_Using_Random_Assignment_Research_Paper.pdf

Kane, T. J., and Staiger, D. O., *Gathering Feedback for Teaching Combining High-Quality Observations with Student Surveys and Achievement Gains*, Bill & Melinda Gates Foundation, January 2012. As of August 14, 2018: https://files.eric.ed.gov/fulltext/ED540960.pdf

Korn, S., Gamboa, M., and Polikoff, M., "Just How Common Are the Standards in Common Core States?" C-SAIL Blog, November 3, 2016. As of August 27, 2018: https://www.c-sail.org/resources/blog/just-how-common-are-standards-common-core-states

Martínez, J. F, Borko, H., and Stecher, B., "Measuring Instructional Practices in Middle School Science Using Classroom Artifacts," *Journal for Research in Science Teaching*, Vol. 41, No. 1, 2011, pp. 38–67.

Measures of Effective Teaching Longitudinal Database, website, undated. As of August 14, 2018: https://www.icpsr.umich.edu/icpsrweb/METLDB/

Mihaly, K., Schwartz, H. L., Opper, I. M., Grimm, G., Rodriguez, L., and Mariano, L., "Impacts of a Checklist on the Feedback Principals Provide to Teachers After Observations," *IES Report*, Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest, REL 2017-285, 2018. As of August 28, 2018: https://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/REL_2018285.pdf

Norton, J., Ash, J., and Ballinger, S, "Common Core Revisions: What Are States Really Changing?" ABT Associates, 2017. As of February 27, 2018: http://abtassociates.com/Perspectives/January-2017/Common-CoreRevisions-What-Are-States-Really-Chang

Pianta, R. C., and Hamre, B. K., "Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation can Leverage Capacity," *Educational Researcher*, Vol. 38, No. 2, 2009, pp. 109–119.

Reform Support Network, *Making High-Quality Teacher Evaluation Manageable*, Washington, D.C., 2013. As of August 27, 2018: https://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/teacher-evaluation-manageable.pdf

Rowan, B., and Correnti, R., "Studying Reading Instruction with Teacher Logs: Lessons from a Study of Instructional Improvement," *Educational Researcher,* Vol. 38, No. 2, 2009, pp. 120–131.

Taylor, E. S., and Tyler, J. H., "The Effect of Evaluation on Teacher Performance," *American Economic Review*, Vol. 102, No. 7, 2012, pp. 3628–3651.

TNTP, *TNTP Core Teaching Rubric: A Tool for Conducting Common Core–Aligned Classroom Observations*, New York, February 18, 2014. As of August 14, 2018: https://tntp.org/publications/view/tntp-core-teaching-rubric-a-tool-for-conducting-classroom-observations

U.S. Government Accountability Office, *Race to the Top: States Implementing Teacher and Principal Evaluation Systems Despite Challenges*, Washington, D.C., GAO-13-777, 2013.

Wiener, R., "Teaching to the Core: Integrating Implementation of Common Core and Teacher Effectiveness Policies," Aspen Institute, 2013. As of August 14, 2018: https://files.eric.ed.gov/fulltext/ED542704.pdf