RAND CORPORATION

RESEARCH BRIEF

Considerations for Implementing the TNTP Core Rubric

TNTP—a nonprofit organization that seeks to improve student educational outcomes and reduce educational inequality by working with high-need schools to attract, train, and retain teachers—designed the TNTP Core Teaching Rubric as a classroom observation tool to be used as part of a multi-measure evaluation system. TNTP Core is unique as a classroom observation tool because it focuses on the activities of students in the classroom, rather than the more typical focus on the teacher. TNTP Core promotes both positive teacher development and high-quality instruction essential to supporting Common Core State Standards in the classroom. TNTP asked RAND Education to assist with the further development of this classroom observation tool by answering two questions about the rubric's fairness, meaningfulness, and reliability:

- What is the relationship between teachers' TNTP Core scores and their students' performance on math and English language arts (ELA) assessments?
- How does raters' content expertise affect the consistency and accuracy of the TNTP Core rubric?

To answer these questions, RAND researchers analyzed trained raters' scorings of video-recorded classroom instruction using TNTP Core. Teachers volunteered to participate in the study and understood that the ratings were for research purposes only and would not affect their formal evaluation process. The researchers also used teachers' math and ELA value-added scores as the comparative measure of instructional quality. Value-added scores captured teachers' contributions to students' math and ELA achievement, net of factors outside of teachers' control. The value-added scores for this study were drawn from a subset of participating teachers.

How TNTP Core Is Different from Other Classroom Observation Protocols

TNTP Core was designed to efficiently describe high-quality, standards-aligned instructional practice. Many observation rubrics are scored based on quantifiable teacher behaviors in the classroom, but TNTP Core focuses instead on student experiences, engagement, and outcomes. This student-focused rubric is made up of the four domains described in Table 1. Each domain is scored on a five-point scale. The concise design of the protocol is intended to help teachers and administrators focus on actionable feedback, and to promote professional development within the context of a multidimensional evaluation system.

Key findings:

- There are modest relationships between teachers' TNTP Core scores and student achievement gains in English language arts. There are no statistically significant relationships between TNTP Core scores and achievement gains in math.
- A teacher's TNTP Core classroom observation protocol score may be influenced by raters' content-area expertise. Schools should have raters from different backgrounds rate different lessons, and schools should have high standards for raters in general.
- Teachers' overall instructional practices may not be fully captured by TNTP Core raters if observations are too infrequent. Schools should consider building internal systems and activities to monitor score quality.

Table 1. Teacher Performance Areas Measured by the TNTP Core Rubric

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

Findings

There are many factors to consider when determining whether classroom observations provide high-quality information for fair and consistent feedback and evaluation. The RAND team focused on three: meaningfulness, fairness, and reliability. This brief highlights key findings related to each of these aspects, and closes with the researchers' recommendations on how to implement the TNTP Core observation rubric.

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

The TNTP Core rubric was designed to assess students' engagement with grade-level appropriate content, and the practices associated with good instruction within each TNTP Core domain are assumed to also affect student achievement. The RAND team assessed whether teachers' TNTP Core scores were associated with students' math and ELA learning gains. The results suggest a modest relationship that differs by subject area and TNTP Core domain. While TNTP Core scores and teachers' value-added scores generally move in the same direction (e.g., teachers with better TNTP Core scores have slightly higher value-added scores than teachers who have lower TNTP Core scores), the modest relationships suggest that there are additional factors besides TNTP Core scores that would explain differences in teachers' value-added scores.

Disagreements between raters introduced uncertainty into TNTP Core scores.

The analysis also compared how often two raters—one a subject area expert and one a generalist—agreed on a teacher's lesson in each of the four domains. Across all four domains and subjects, raters often disagreed on their judgments about the quality of a lesson. Overall, agreement rates were roughly equal to what would be expected by chance. For math, raters' content knowledge not only influenced the agreement between two raters; math specialists (math raters with strong math content knowledge) also tended to give lower ratings than raters with more generalist training. The full report examines the implications of this and offers recommendations for reducing the impact of rater disagreement when using TNTP Core.

Reliable measurement of teaching quality from the TNTP Core rubric requires many observations by many raters.

The researchers also assessed the reliability of TNTP Core scores. The intuition for the analysis is to ask how many observations would need to be conducted to get an accurate measure of teachers' actual instructional practices. If teachers' TNTP Core scores differ substantially across lessons and raters, then it will take a large number of observations to make general claims about a teachers' instructional quality. The results suggest that a large number of observations are necessary to reach benchmark levels of consistency put forth by the education and psychological research community and recommended for the general use of observational protocols in teacher evaluation. School systems implementing TNTP Core will have to balance the number of observations suggested by the study's results with the budget and staff time constraints. To make general claims about a teacher's instructional quality from TNTP Core scores, and observational protocols in general, one would need many observations of a given teacher conducted by many raters throughout the year. The results suggest that even when teachers were observed four times a year each time by two raters, there was still considerable uncertainty about the quality of teachers' instructional practice. For example, if a teacher was observed four times by two raters and received an average score of 3 for the Academic Ownership domain, the uncertainty in the rubric means that a teacher's true score could be between 2.3 and 3.7.

The uncertainty affects the use of the TNTP Core rubric in two ways. First, the added noise makes it more challenging to measure the relationship between TNTP Core scores and other measures of teacher quality (e.g., value-added scores). Second, the uncertainty can diminish the usefulness of feedback that can be given if it is based on TNTP Core scores. If TNTP Core scores do not accurately represent the overall practices of a teacher, then the utility of the scores for diagnosing and developing instructional practices is compromised.

It is important to keep two considerations in mind with reliability findings in the study and the use of TNTP Core. First, this study used individual lessons as the unit of analysis, which will likely have lower reliability than aggregating TNTP Core scores to grade, school, or district levels. Second, to ease the implementation of the research project, the researchers gathered lesson data from video-recording teachers, and raters watched these video-recorded lessons to generate TNTP Core scores. In practice, however, most school systems would rely on in-person observations. Evidence from the field suggests that there may be differences in the reliability of video versus in-person ratings. These are both areas of further study on TNTP Core.

Recommendations

Classroom observation protocols are difficult to implement well, and they rely on a number of factors, from rubric design to training raters to the complexity of teaching. It is not surprising that uncertainty is introduced in a number of ways throughout implementation of the TNTP Core rubric (and most observational protocols). Yet there are several actions that school, school district, and charter network leaders can take to improve the accuracy and consistency of scores from the TNTP Core rubric.

Set high standards for rater certification.

Raters play a crucial role in the implementation of any observational protocol. To obtain useful information from the rubric, it is necessary to use highly trained raters. Raters need to become experts in TNTP Core's four dimensions of teacher practice and how to accurately apply this knowledge to specific ratings. TNTP hired generalist and specialist raters to participate in this study. Before reviewing videos, reviewers went through TNTP's norming process, which includes participating in virtual training and extensive independent study, guided practice, and webinar-based assessment and feedback. The reviewers then rated at least five lesson videos on which TNTP assessed their alignment to the normed ratings, receiving feedback via group webinar after each lesson. In order to pass the training, reviewers had to meet TNTP's 75/50 threshold—meaning that, on average across all five lesson videos, the rater must be within one rating of the normed rating up or down on three of four performance areas (75 percent), and at least two of the five ratings must be an exact match with the normed rating (50 percent). The full report provides more detail on TNTP's norming process.

The report also provides an overview of good rater certification practices for the use of observational protocols to measure instructional quality. Raters need to become experts in the measure's domains (e.g., TNTP Core's four dimensions of teacher practice) and how to accurately apply this knowledge to specific ratings. Typically, when implementing a protocol, raters need to go through a certification process to ensure that they would score a lesson similar to an expert rater. There are a number of ways to certify raters. Ideally, the certification process would include three key aspects. First, during the training period, raters need access to a set of lessons with known scores from master raters. This way, raters' assessment of instructional quality can be compared against known scores from master raters with deep expertise in the tool. Second, raters need to hit a high level of agreement against these master raters' scores. While a hard and fast rule does not exist in the literature, ideally raters would reach high levels of exact agreement and one-off agreement (e.g., rater trainee and master rater scores are off by only one point). For example, raters could have to meet 75 percent exact agreement and 90 percent one-off agreement across five lessons in order to ensure mastery for the rater trainings.

Third, raters need to be recertified on a continuous basis. While initial mastery is important, raters tend to drift from their original certification standards as time elapses from their training. For example, it would be useful for schools and school districts to embed frequent post-certification calibration and validation exercises during any period in which a rater is providing teachers' feedback through the TNTP Core rubric. Calibration and validation are similar ideas related to ensuring a high level of accuracy for raters. Calibration is the same process raters go through during the original certification process: watching and scoring lessons previously rated by a master rater. Validation is a process whereby master raters would watch a subset of lessons assigned to a given rater. In both cases, master raters and raters would discuss score discrepancies.

Ensure equitable and diverse spread of raters across lessons. The study suggests that raters vary in their content knowledge and that content expertise may influence the severity or leniency of teacher observation ratings. However, the study was not designed to evaluate whether content experts or generalists more accurately measured instructional quality using TNTP Core—for example, the researchers did not have known TNTP Core scores from each lesson from master raters. To promote fair implementation of observational rubrics, school leaders should use both types of raters and ensure that they are spread evenly across teachers and lessons. One way to do this would be to administer a content knowledge assessment to all prospective raters, or to use some other measure of raters' content knowledge, and ensure that rater content knowledge is spread evenly across teachers and lessons.

Consider collecting additional evidence of teacher quality when using TNTP Core to measure instructional quality. TNTP Core was designed to measure one facet of a teacher's instructional practice through observations, by focusing on standards-aligned instruction and student activities. The results of this study are consistent with the broader recommendation that, to best assess and support teachers' instructional practice and professional development, school systems should collect multiple measures of teacher and instructional quality. Accuracy and consistency are not static properties of observation protocols—they depend on who is observed and the conditions under which scores are collected. Scores collected by researchers or external observers may have different properties than scores collected from principals or peers. Additionally, it is crucial to think about accuracy and consistency along with how the rubric will be used. Scores could potentially be used in a wide range of ways: to promote grade-level or content-area conversations, to guide or inform professional development plans, or to make performance evaluation decisions. Different kinds of evidence would be necessary to support these intended uses; uses that are tied to consequences require a strong evidence base and a lower tolerance for uncertainty. As schools, 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. If TNTP Core is used as one measure in a multiple-measure system, separate evidence should be collected supporting inferences based on multiple measures.

This report was commissioned by TNTP with the support of the Charles and Lynn Schusterman Family Foundation.

This brief describes research conducted in RAND Education and Labor and documented in *Validation Study of the TNTP Core Teaching Rubric*, by Andrew McEachin, Jonathan Schweig, Rachel Perera, and Isaac M. Opper, RR-2623-NTP, 2018 (available at www.rand.org/t/RR2623). To view this brief online, visit www.rand.org/t/RB10033. The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors. **RAND**® is a registered trademark. © RAND 2018

www.rand.org