Redefining the Instructional Leader: Principals' Use of Process and Outcome Measures of Teacher Quality

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citation</th>
<th>Holcombe, Rebecca Wood. 2016. Redefining the Instructional Leader: Principals' Use of Process and Outcome Measures of Teacher Quality. Doctoral dissertation, Harvard Graduate School of Education.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.Inst">http://nrs.harvard.edu/urn-3:HUL.Inst</a> Repos:27112689</td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.Inst">http://nrs.harvard.edu/urn-3:HUL.Inst</a> Repos:dash.current.terms-of-use#LAA</td>
</tr>
</tbody>
</table>
Redefining the Instructional Leader:  
Principals' Use of Process and Outcome Measures of Teacher Quality

Rebecca Holcombe

Susan Moore Johnson  
Sarah Dryden-Peterson  
Monica Higgins

A Thesis Presented to the Faculty  
of the Graduate School of Education of Harvard University  
in Partial Fulfillment of the Requirements  
for the Degree of Doctor of Education

December 6, 2015
Acknowledgements

This work reflects the faith and support of so many people, who pushed me to finish when I was content to move on. They were right, of course, and I am deeply grateful to them for telling me to just get it done. Any acknowledgement will be an understatement of their support.

First, I thank the principals who gave me their time, and without whom this work could not have been done. I am humbled by the work they are doing and the wisdom they bring to their schools, and know that nothing is more important than showing up, every day, in schools, to use that wisdom to create better opportunities for every one of our children. I also thank Noah Bookman and Donna Muncey of LAUSD, who gave me access to the principals, were generous with their time, and are working hard to support improvement. In particular, the careful planning and support around the TLF appeared to have a powerful and positive impact on conversations about teaching and learning in most of the schools I visited. I have been around a few schools, and can’t think of any other intervention at scale that I have seen that had this kind of impact on professional conversations. I hope the energy and quality of implementation can be sustained as this effort scales up, and I look forward to learning from this experiment.

Second, I thank my advisors and mentors at Harvard, who kept me going. Thanks to John Willett, who helped me at a critical transition and who changed the way I thought about data. Thank you also to Monica Higgins and Sarah Dryden-Peterson, whose careful and detailed and specific feedback helped me beat this into its current shape. I give a special thanks to Susan Moore Johnson, whose commitment to educators and practice has
inspired me for years, and who believed in this work when I was ready to drop it in the
“could have . . .” file and stay in the world of practice. Her careful reading and feedback
guided my efforts to bring this dissertation into being.

Thanks to Gregory Foster for his careful proofreading.

Thank you to my dear friend, statistics partner, and professional colleague Amy
Fowler, with her unsurpassed wisdom, insight, and sharp humor. Amy believes, as I do,
that we persevere in what we do because every child deserves the best education we can
give her or him. Amy keeps me on the rails and reminds me of what matters most in
schools, every time.

Many members of my family also share in this. I am privileged to have parents
who expected a lot of me, and also were able to offer support when I needed it. My dad
modeled relentless pursuit of goals, and my mother is a mentor and source of inspiration,
who shares my belief that there is a policy tension between evaluation and development
of human capability. She challenged me, engaged with my ideas, and read and provided
feedback on this work. On several occasions, she cared for my children, so I could write.
Coincidentally, we selected the same Hertzberg quote to frame our dissertations. Her own
life journey is a tribute to the power and privilege of education, and reminds me to use
this power I now have to make sure others have the same opportunity.

I thank also the Finers, who are themselves a village that continues to help raise
several children, including me and my own, and the Millikens and Burnhams, with whom
I now hope to take on my share of the carpooling.

Balancing work, doctoral programs, and life is a challenge that has not always
been pretty, and I thank my children for tolerating my distraction and reminding me
every day of what miracles human beings are and how important each child is. You inspire me and make me proud.

Most of all, I thank my husband James, who is my partner in everything that is worthwhile, and without whose support this would never have been finished. I am blessed with a smart, funny, supportive partner who tolerates my eccentricities, makes tea, takes care of the children, and tells me to calm down and get cracking, all while getting on with his own work. We should all be so lucky.
# Table of Contents

**ACKNOWLEDGEMENTS** ............................................................................................................................................. I  
**ABSTRACT** ..................................................................................................................................................................... VI  

## CHAPTER 1: INTRODUCTION ........................................................................................................................................... 1  
- The Changing Nature of Teacher Evaluation ............................................................................................................. 1  
- Criteria for Evaluating the Quality of Measures of Teacher Efficacy ........................................................................ 9  
- The Relationship between Rankings Based on Gain Measures and Principal Observations .................................... 13  
- Teacher Responses to Testing under High-Stakes Conditions and the Implications for the Validity of Inferences about Teacher Quality Based on Student Test Scores .................................................. 14  
- The Principal’s Role with Respect to Teacher Evaluation ....................................................................................... 16  
- Current Study ......................................................................................................................................................... 18  

## CHAPTER 2: CONTEXT ....................................................................................................................................................... 22  
- Los Angeles and the Educator Growth and Development Program .............................................................................. 22  
- Academic Growth over Time Rankings ..................................................................................................................... 22  
- The Teaching and Learning Framework .................................................................................................................... 24  
- Reconciliation of the AGT and the TLF ...................................................................................................................... 27  
- Accountability Context ................................................................................................................................................... 28  
- Proficiency Thresholds and “Percent of Students Scoring as Proficient” (Federal Accountability) ........................................... 28  
- Academic Performance Index (State Accountability) ................................................................................................ 29  
- School Performance Framework (LAUSD Accountability) ............................................................................................. 30  

## CHAPTER 3: RESEARCH METHODS .................................................................................................................................. 32  
- Sample ........................................................................................................................................................................... 32  
- Data ............................................................................................................................................................................... 38  
- Data Collection and Audit Trail ..................................................................................................................................... 41  
- Data Analysis .............................................................................................................................................................. 42  
- Validity ............................................................................................................................................................................. 43  

## CHAPTER 4: WHAT DID PRINCIPALS THINK THE AGT RANKING ACTUALLY MEANS? ............................................................... 46  
- The AGT as a Measure of Teacher Quality .................................................................................................................. 46  
- Factors Principals Believe Shape the Accuracy and Reliability of the AGT Rankings ................................................. 49  
- The AGT as a Relative Measure or an Absolute Measure? What Principals Want ..................................................... 74  

## CHAPTER 5: THE AGT AS A LEVER TO IMPROVE PRACTICE OR A HAMMER THAT ERODES PRACTICE? ......................... 78  
- Using the AGT to Identify Strengths and Weaknesses of Teachers and Target Improvement Efforts ................................. 80  
- The AGT as a Lever to Improve Practice ..................................................................................................................... 82  
- Use of the AGT by Principals in Lower-Scoring, Higher-Poverty Schools and in Higher-Scoring, Lower-Poverty Schools .......................................................................................................................... 85  
- Aligning Instruction to the Test ................................................................................................................................... 97  

## CHAPTER 6: WHAT DO PRINCIPALS THINK IS MEASURED BY STANDARDS-BASED OBSERVATIONS BASED ON THE TEACHING AND LEARNING FRAMEWORK? ................. 105  
- The Teaching and Learning Framework as a Measure ................................................................................................. 106  
- Broadening of Principals’ Operational Understanding of Effective Teaching ............................................................ 110  
- Factors that Affect the Quality and Accuracy of TLF Ratings .................................................................................. 113
Abstract

Recent innovations in teacher evaluation include the development of standards-based frameworks for observational evaluations of teacher practice and development of test-based measures of teacher effectiveness. These tools raise questions about how best to identify good teaching, about the roles of both evaluations and principals in improving teaching, and about the relationship between school context and teacher evaluation.

This qualitative study examines principals’ understanding and use of a test-based ranking of teacher quality—the Academic Growth over Time (AGT) ranking—and a standards-based observation framework—the Teaching and Learning Framework (TLF)—in the Los Angeles Unified School District.

Findings suggest principals believed outcomes matter, but harbored concerns about what AGT rankings actually measured. In addition, principals felt AGT rankings provided little information about how to improve teaching. Some were concerned that in low-scoring schools, the AGT rankings might reinforce test-focused instruction in ways that disadvantaged students.

Principals preferred the TLF because it was more comprehensive than previous observation tools, helped teachers evaluate their practice against clearly defined expectations, and helped principals improve instruction. Though principals used the TLF to assign teachers ratings, they primarily described it as a tool to improve the quality of teaching in their buildings. Principals noted that the TLF process placed heavy demands on the time of teachers and principals, and expressed concern about how they could sustain high-quality implementation of the TLF when they had to use it at scale.
Most principals noted cases where the AGT rankings indicated a different reality about a teacher’s effectiveness than what the principals expected, based on classroom observations. How principals reconciled discrepancies varied. Principals in higher-scoring schools were more likely to discount test-based measures and more likely to emphasize the professional capability of their staff as an asset.

The findings suggest principals’ use of evaluation tools is mediated by their confidence in the tools and their own leadership, by how useful they feel the tools are, and by their own school contexts. In addition, this study suggests that the unique burdens of persistent socioeconomic and racial segregation may hamper local efforts to implement promising practices at some sites.
“We shall never understand the reality of organizations if we persist in studying them from a distance, in large samples, with gross cross-sectional measures. We learn how birds fly by studying them one at a time, not by scanning flocks of them on radar screens.”

-Mintzberg, The Nature of Managerial Work, 1973

Chapter 1: Introduction

The Changing Nature of Teacher Evaluation

Historically, principals have been the sole arbiters of teacher quality, and evaluations based on principal observations were the basis for decisions regarding retention or promotion. These observations were “process” measures: typically, they assess the extent to which observed teaching practices conform to established expectations. Principals observed the practice of teaching, and based on what they saw, gave teachers an assessment. For years, the authority of principals to make these judgments received little scrutiny.

The No Child Left Behind Act (NCLB) ushered in new emphasis on outcome measures of school effectiveness, defined in terms of test-based measures of external accountability for learning (Slavin, 2002; Wong & Nicotera, 2007). Many researchers and policymakers criticized evaluations based solely on observations, both for failing to distinguish between high- and low-quality teachers and for failing to take into account improvements in student learning over the academic year immediately preceding the observation period (Glazerman et al., 2010; Weisberg, Sexton, Mulhern, & Keeling, 2009). These constituencies also note that principal observations are vulnerable to observer bias, and if not calibrated, cannot support valid inferences about teacher quality.
During the era of test-based accountability ushered in by NCLB, new studies suggested test-based measures offer a more objective assessment of teacher quality (Gordon, Kane, & Staiger, 2006). Since then, a growing body of work has examined the quality of test-based value-added measures of teacher quality, as well as differences in identification of quality by different models using the same data (see, e.g., Ballou, 2009; Harris, 2009; Koedel & Betts, 2010; 2011; McCaffrey, Koretz, Lockwood, & Hamilton, 2004; Papay, 2011).

More recently, the “Race to the Top” initiative required applicants to prioritize “recruiting, developing, rewarding, and retaining effective teachers and principals, especially where they are needed most,” but more importantly, defined “highly effective teachers” as teachers “whose students achieve high rates (e.g., one and one-half grade levels in an academic year) of student growth” (Federal Register, 2009, p. 59839). Similarly, highly effective principals are defined as principals whose students “overall and for each subgroup, achieve high rates (e.g., one and one-half grade levels in an academic year) of student growth” (Federal Register, 2009, p. 59839). Because Race to the Top requires principals and teachers to be evaluated in "significant part" (Federal Register, 2009, p. 59839) based on changes in student achievement on state tests used for accountability purposes and to use these data for employment decisions, both principals and teachers are highly incentivized to strive for large student gains on tests used for accountability purposes.

Given this increased reliance on test-based measures of student learning and teacher effectiveness, the authority of principals with respect to teacher evaluation is now subject to increasing scrutiny. Even though principal observations appear to be
commonly used for teacher evaluations (Brandt et al., 2007), recent research argues there is considerable variability in teacher quality (Aaronson, Barrow, & Sander, 2003; Kane, Rockoff, & Staiger, 2008; Leithwood, Louis, Anderson, & Wahlstrom, 2004; Rivkin, Hanushek, & Kain, 2005; OECD, 2005; Rockoff, 2004), and that traditional observation-based models of teacher evaluation fail to distinguish between high-quality and low-quality teachers and to take into account improvements in student learning (Glazerman et al., 2010; Weisberg et al., 2009). Because teacher quality is now viewed as the most important driver of student learning over which districts have control (Aaronson et al., 2003; Hanushek & Rivkin, 2010; Kane & Staiger, 2008; Kane, Rockoff, & Staiger, 2008; Nye, Konstantopoulos, & Hedges, 2004; Rivkin et al., 2005; Rockoff, 2004), districts experience pressure to develop better tools for identifying and rewarding teachers whose students demonstrate greater gains on tests used for accountability purposes.

As a result, many states and school districts, including the Los Angeles Unified School District (LAUSD), introduced models for evaluating teachers based on a combination of principal evaluations and value-added measures that account for student learning gains. As Noah Bookman, a policy advisor for LAUSD, told the Los Angeles Daily News, “Academic growth over time (AGT) provides us with an apples-to-apples comparison of how well schools and educators are doing with taking their students from point A to point B” (Llanos, 2011). These models promise to provide a more objective measure of teacher quality (Glazerman et al., 2010; Kane & Staiger, 2012) and are viewed by many as a more reliable and less biased alternative to principals’ observations (Gordon, Kane, & Staiger, 2006). Research exploring the relationship between student gain scores and observational measures of effective teaching suggests that teachers whose
students show greater gains also tend to receive higher ratings based on observations (Kane & Staiger, 2012; Stronge, Ward, Tucker, & Hindman, 2008).

However, there is also substantial evidence that most tests are predictably imperfect samples of the domains they sample, and thus teaching to the tests is not the same as teaching to the standards (Holcombe, 2012). Under high-stakes conditions, such as when teacher evaluations or school status are dependent on test scores, educators may face strong incentives to focus narrowly on evidence that drives test scores, or to otherwise teach in ways that raise scores on the tests used for accountability purposes, but which do not lead to generalizable improvements in student learning.

Increased reliance on value-added measures to evaluate teachers potentially alters the role of the principal and potentially shapes principals’ professional judgment in ways that are not yet fully understood. If value-added scores are perceived as being more objective than ratings based on observations, will this change principals’ sense of their authority as the arbiters of teacher quality in their schools? Will some principals modify their observation-based ratings of teacher quality in response to test-based measures of teacher quality, and if so, how? How do principals understand the evidence when a principal appears to be rating teachers more harshly or more leniently than might be expected, given value-added scores? Do some principals privilege dimensions of teacher quality that are not captured by student gain scores or question what is measured by scores?

Currently, no studies have gone beyond correlations of rankings based on observations and rankings based on gain scores to explore underlying mechanisms that explain observed patterns in correlations, or how principals understand these patterns.
What do principals think each of these measures (observational and test-based) actually measures, and how might this explain the relatively low correlations between the two? Moreover, no study examines: (a) how principals make sense of the relationship between the ratings they and their peers assign, and test-based rankings of teacher efficacy; or (b) how principals perceive the effect of increasing emphasis on value-added measures on their role and authority.

This study starts from the premise that we cannot fully understand how and why principals use test-based measures or standards-based observations to evaluate teachers and improve learning by examining large-scale data on implementation or cross-sectional panels of test scores. Instead, to paraphrase Mintzberg (1973), we learn how principals use evaluation tools to improve learning by studying them directly, one at a time in their own context. By understanding principal perceptions of the value of different evaluation tools, as well as the potential challenges to successful implementation across a diversity of contexts, we can gain insight into why promising practices may or may not be successfully implemented or yield the intended outcomes.

A modest but solid body of research suggests that some principals are skeptical of the test-based models, and more likely to trust ratings based on observations of practice using standards-based frameworks. In interviews with school leaders, Goldring and others (2015) identified numerous concerns held about teacher evaluation tools based on student test scores, including perceptions of problems related to the timing of the release of these measures, the validity of these measures, the utility of these measures with respect to identifying the mechanisms by which teachers can improve practice, and a lack of clarity with how the test score–based models actually work. In part as a result of these
concerns, principals in Goldring and others’ study indicated that increasingly, they are focusing on standards-based teacher observations in their decision-making. The study also identified three additional benefits to the standards-based observations that led principals to privilege them as tools for measuring efficacy and supporting improvements in practice: “(1) the consistency of the data, (2) the transparency of the data, and (3) the specificity of evidence of teacher observation data” (Goldring et al., 2015, p. 97).

The current study builds on this work by studying how principals in the pilot phase of LAUSD’s Educator Growth and Development Cycle used a test-based measure of student gains associated with individual teachers—Academic Growth over Time (AGT)—and a standards-based observation protocol—the Teaching and Learning Framework (TLF)—to evaluate teachers and support improvements in teacher practice. In theory, if the teacher behaviors described in the TLF are truly effective, then that effectiveness will be picked up in the form of improved student performance on standardized tests, and in turn, AGT rankings and TLF ratings should be correlated. Several obvious factors may attenuate the correlation; notably, the AGT is a relative ranking, and the TLF rubric provides a standards-based rating. Still unexplored are questions related to what factors principals perceive might affect the accuracy and reliability of their observations, as well as the correlation between the two measures. In addition, this study explores the relationship between the AGT and the TLF in principals’ professional judgments about teacher quality.

Beyond the goal of measuring teacher quality, standards-based observation tools are also increasingly used to support improvements in teacher quality. This suggests an evolution in the purpose of evaluation. The specificity in the standards-based frameworks,
including the rubrics that include descriptions of what practice looks like at different levels of teacher effectiveness, potentially lends itself to formative conversations and a focus on growth and improvement in teaching practice. As Goldring and others (2015) noted, the standardized observation process provides a “bigger picture” of the teacher’s performance, it can inform individualized and large group professional development, and it forms the basis of individualized support for remediation plans that serve as the documentation for dismissal cases. It helps principals provides specific and ongoing feedback to teachers. (p. 102)

In other words, one of the primary appeals of these frameworks is their ability to provide support for continuous growth and improvement in practice. As Elmore (2005) has noted elsewhere, student scores may not increase in response to higher stakes for teachers if the reason teachers are not teaching more effectively is that they do not know how to do so.

The promise of standards-based frameworks for evaluation and feedback is that they provide teachers with the means to improve instruction, on the theory that in turn, the scores will take care of themselves.

This raises questions about the relationship between evaluation and improvement of practice in the minds of principals. Importantly, is their job to monitor quality or to improve it? If both, which tools better serve each purpose? Some researchers have suggested that the relationship between high-stakes accountability and learning is complicated, and that heavy emphasis on accountability and evaluation can in fact potentially undermine efforts to improve practice (Murphy et al., 2014). Papay (2012) noted that most policy debates have focused on the first purpose—measuring teacher
quality and holding them accountable for performance—but that in truth, the second purpose—highlighting strengths and weakness and supporting continued professional development—is equally if not more important. Given the challenges of recruitment and retention, particularly in high-poverty contexts, this latter purpose is potentially a more promising way to invest in improving the instruction to which all children have access.

In addition, the measures of teacher quality are only useful to the extent that those tasked with using them understand and can use them well. The current study draws on recent research that argues that, instead of thinking of implementation challenges as purely technical (e.g., poor communication of policy goals or a mismatch in goals between policy makers and those tasked with implementation), we also need to understand how implementers understand and make sense of policy and tools. Spillane and others stated, “From a cognitive perspective, (effective) implementation hinges on whether and in what ways local implementing agents understand how policy demands impact the extent to which they reinforce or alter their practice” (Honig, 2006, p. 47). For example, policymakers may think value-added measures are important, objective tools for the purpose of teacher evaluation, but if principals doubt or distrust test-based rankings, they are unlikely to use them in practice in the way policy makers intended.

The principal’s role in instructional leadership has traditionally been thought of as one of communicating high expectations for teachers and students, supervising instruction, monitoring assessment and student progress, coordinating the school’s curriculum, promoting a climate for learning, and creating a supportive work environment (Marks & Printy, 2003; Murphy, 1990). In contrast, more recently, principal instructional leadership has shifted from a focus on the principal as “an inspector of
“teacher competence” to the principal as “a facilitator of teacher growth” (Marks & Printy, 2003, p. 374). Collaborative inquiry with teachers, creating opportunities for reflection, discourse, and professional growth, and the development of professional learning communities have all been part of this shift (Dufour & Eaker, 1998; Kruse, Louis, & Bryk, 1995; Marks & Printy, 2003). To facilitate professional growth, principals need to not only understand the tools at their disposal, but they also need to exercise judgment about when and how to use them. This invites several questions, which will be explored in this study, including:

- Which measures of teacher quality best support principals’ efforts to facilitate instructional improvement?
- Does the context in which a principal operates shape which tools he or she can use effectively?

**Criteria for Evaluating the Quality of Measures of Teacher Efficacy**

Papay (2012) stated, “As measurement tools, teacher evaluations—both value-added and standards-based observations—should be judged according to three criteria: are they unbiased, reliable and valid?” Papay provided an excellent and comprehensive discussion of the threats to validity associated with both test-based and observational measures, and he concludes that both test-based and observational evaluations are subject to several weaknesses that threaten the validity of inferences about quality based on those measures.

For example, principals may bring preconceptions about a teacher to observations, and these preconceptions may in turn bias their ratings. In addition, if different principals have different standards, we cannot be certain that the rating assigned to a teacher by one
principal is the same rating that a different principal would have assigned if completing the same observation using the same tool. Evidence from several sources suggests that sufficient training on a quality framework and follow-up calibration can enhance the reliability and accuracy of ratings by principals, but not guarantee the elimination of bias (Fowler, 2013). Fowler (2013) found that some principals, for example, tend to be harsh or lenient as raters. It is unclear whether, even after calibration, these principals stay calibrated or drift back to being harsh or lenient raters. Overall, however, what research is available suggests that with appropriate training and a focus on evidence related to clearly specified standards, the accuracy and consistency of principal ratings can be improved (Fowler, 2013).

In contrast, test-based measures have the appearance of being objective, because they are quantifiable and generated from ostensibly objective standardized test scores. However, in recent years, extensive modeling and analysis have suggested that these measures might not adequately account for all variables that explain student performance and also may be quite sensitive to a variety of factors, including test scaling, non-random student sorting, and ceiling effects (see, e.g., Ballou, 2009; Briggs & Weeks, 2011; Koedel & Betts, 2010; 2011; Papay, 2011). These factors would systematically bias the test-based rankings of at least some teachers.

Field trials of the effects of different treatments typically involve a handful of treatments and hundreds of subjects assigned to each treatment. Thus, when different treatments exhibit different effects, we can be reasonably confident that the differences in effect reflect real differences associated with the different treatments. As Braun (2006) pointed out, however, test-based measures of teacher quality treat each teacher as a
separate treatment, and assume classes are randomly assigned and roughly equivalent. However, if student assignment is non-random, as it often is, this would bias estimates.

Relevant to the current study is the question of whether using test-based tools and observation-based tools in tandem might introduce bias into observational judgments. For example, the AGT rankings provide an external, apparently “objective” judgment of the effectiveness of teachers relative to other teachers in the district. When a principal’s rating of a teacher using the TLF tells a different story than might be expected, given the teacher’s associated AGT ranking, how does the principal make sense of that discrepancy? Evidence on this type of scenario is scant, but Jacob and Lefgren (2008) found evidence that the test-based accountability model in which principals and teachers operate may influence their assessments of teacher effectiveness. Specifically, Jacob and Lefgren noted that principal observations may be biased by two phenomena: (a) they may be more focused on low-performing, as opposed to high-performing students, and (b) they may be focused on achievement levels (as defined, perhaps, by NCLB) as opposed to how much value a teacher adds to the students’ learning. In other words, how principals operationalize their assessments of teacher quality may in part be influenced by how test-based accountability policies define teacher quality.

An additional threat to the validity of inferences about teacher quality based on principal observations is reliability. A measure is reliable if it captures true underlying performance, such that the rating or ranking on one day is the same the teacher would receive on a different day, with a different lesson or test, or if observed by a different principal. Papay (2012) suggested that repeated observations might yield more reliable observation ratings, much as larger testing groups yield more reliable test-based estimates.
In other words, if a principal’s rating of a teacher is based on multiple instances of observation, we may have more confidence the rating is representative of the teacher’s quality. However, principal observations and pre- and post-conferences are time-consuming, and there may be practical limitations on the number of formal observations a principal can conduct.

Issues of reliability, bias, and accuracy all relate to the quality of a teacher evaluation tool as a measure of quality, and specifically, whether the tool can accurately and consistently capture evidence of quality. However, a second purpose of teacher evaluation is to support improvements in teaching practice. Thus, another way to measure the quality of a teacher evaluation tool is in terms of its utility for promoting instructional improvement. Research by Taylor and Tyler (2011) suggests that strong evaluation programs focused on observations and feedback improve teacher effectiveness and, in turn, student achievement as measured on tests. The logic is that if teachers are more effective, students will learn more. To help students learn more, we need to help teachers become more effective.

In this theoretical frame, one indicator that a teacher evaluation system is “good” is the impact it has on improving teacher effectiveness. In this respect, test-based tools and observation-based tools are potentially quite different. One set (test-based tools) are essentially outcome measures that purport to capture the impact of a year of teaching on student learning. These speak to how much “learning” was added, but not to what drove differences in learning across the student population. Teachers may or may not understand the mathematical formula that generates their rankings or how to use them to make changes to their teaching practice.
In contrast, standards-based observation rubrics define a set of preferred instructional practices and student responses (e.g., a definition of what good teaching looks like), and then require principals or observers to audit observed practices and student responses against that description. Observations can be completed at several times throughout the year, and once scripted evidence from the evaluation is matched to the framework, the resulting evaluation feedback provides teachers with a written description of what teaching behaviors were expected, which behaviors were observed, and what “rating” corresponds to the teaching behaviors that teachers demonstrated during the observational process. These data also provide teachers with information on specific changes to practice they can make to improve their rating. Thus, the observational feedback speaks to what teachers can do to improve. In the absence of feedback on practices, teachers may look to the test to identify strategies to improve scores, resulting in test-focused instruction that may or may not actually leave students prepared for more advanced work in subsequent years.

The Relationship between Rankings Based on Gain Measures and Principal Observations

Prior research on the relationship between observational assessments of effective teaching and measures of student achievement and learning suggests that teachers whose students show greater academic growth over time (as measured by performance on standardized tests) also tend to receive higher ratings based on principal observations (Kane & Staiger, 2012; Stronge et al., 2007); however, this relationship is modest. For example, the overall correlation between the two in the LAUSD pilot has been small to
moderate (0.2–0.3) (Bookman, 2012). To the extent that ratings based on observations and rankings based on test-based measures appear to suggest true differences in teacher quality, practitioners may question the validity of judgments about teacher quality based on one, the other, or both of the types of teacher evaluation tools.

Research also suggests that site-specific factors that are not unique to a teacher may explain or influence a given teacher’s associated value-added estimate. For one example, work by Jackson and Bruegmann (2009) suggests that a teacher’s own performance is affected by the quality of her peers, and specifically, that less experienced teachers and teachers who are certified and carry a regular license are more likely to have students who demonstrate score gains when they work alongside strong peers whose students demonstrate score gains. Given that inexperienced and alternatively certified teachers are more likely to be found in higher-poverty schools, this potentially influences AGT rankings differentially across schools. In a district as large as LAUSD, there is considerable variability in circumstances across schools within the district. While site-specific factors that might influence achievement are not evident to those who generate value-added estimates, principals are uniquely positioned to observe—and possibly be influenced by—their perception of these site-level factors.

**Teacher Responses to Testing under High-Stakes Conditions and the Implications for the Validity of Inferences about Teacher Quality Based on Student Test Scores**

The theory of action behind high-stakes testing and accountability is that schools and teachers will be motivated to teach in ways that increase student learning. To evaluate the effectiveness of high-stakes testing as a lever for school reform, we need to
evaluate evidence suggesting that the tests lead to improvements in instruction that, in turn, lead to generalizable improvements in learning. As we do so, our assumption is that test scores are good measures of student mastery of whatever body of knowledge and skills the test samples (Koretz & McCaffrey, 2005; Madaus, 1988), and that score gains indicate generalizable improvements in learning.

However, previous research on test design suggests that the test design process creates several types of opportunities for narrow test-focused instruction that would lead to test gains, but not commensurate gains in student learning (Holcombe, 2012; Holcombe, Jennings, & Koretz, 2013; Koretz and Béguin, 2010). Jennings and Bearak (2010) examined test item–level data from three states, and found that students performed better on the most frequently assessed standards within the test in math and English—a pattern which suggests that at least some teachers may be directing instructional emphasis at material that accounts for the greatest proportion of test points.

If what is tested differs in predictable and substantive ways from what is specified in the state standards, and if teachers use test data to focus or target instruction, the validity of inferences about student mastery of the state standards based on student test scores is threatened. In such a case, we could no longer legitimately claim that rising scores mean that students are learning more or that closing gaps mean greater equity (Haladyna & Downing, 2004; Koretz & McCaffrey, 2005; Mehrens & Kaminski, 1989). In fact, a substantial body of literature suggests teachers respond to high stakes by mirroring the content and appearance of high-stakes tests in their instruction, a practice that potentially inflates scores (Au, 2007; Abrams, Pedulla, & Madaus, 2003; Croft et al., 2005; Hamilton et al., 2007; Herman & Golan, 1993; Jacob, 2005; Jones et al., 1999; Lai
And studies in several states contrasting gains on state tests and the National Assessment of Educational Progress suggest that while gains on NAEP have been relatively flat and performance gaps between different groups of students persist, trends in state scores imply substantial gains and narrowing gaps. Patterns like these suggest that in some classrooms, some portion of observed gains may reflect test-specific knowledge that does not equate with meaningful learning.

Practically speaking, to the extent that any score inflation associated with teaching to the test is uneven across classrooms, test-based measures of teacher quality may lead to inaccurate rankings of teachers with respect to actual teacher quality. In addition, if test emphasis is not well aligned with curriculum standards, test-based measures may misidentify good teachers: teachers who teach the curriculum may not rank as highly as comparable teachers who emphasize tested content. For example, if a science test emphasizes content knowledge, it will reward teachers who emphasize content knowledge, even if the emphasis in the curriculum frameworks is on developing student inquiry.

The Principal’s Role with Respect to Teacher Evaluation

The strong policy importance placed on improving instruction has changed the role of principals. On the one hand, principals are leaders of instruction, and one of their more important responsibilities is achieving consistent improvement of instruction. In this frame, the purpose of evaluation and standards-based protocols is to support
improvement of teaching. Principals using these protocols spend substantial blocks of time in classrooms, meeting with teachers, providing feedback, and fostering a shared vision of quality instruction. In the report on initial findings from the Measures of Effective Teaching Project (Bill and Melinda Gates Foundation, 2010), the authors suggested that effective observation tools, in an ideal world, provide for a shared language of practice and a shared understanding of effective teaching. Research suggests that feedback and ongoing professional development that target an individual teacher’s specific strengths and weaknesses over time lead to better practice, as identified through structured observations and scripting of evidence by principals (Bill and Melinda Gates Foundation, 2010; Kane et al., 2013; Tyler & Taylor, 2011). By shaping how teachers discuss teaching and providing targeted feedback, principals can have a strong—if indirect—effect on how teachers make sense of policies and how performance improvement efforts are implemented in practice (Coburn, 2001; Marks & Printy, 2003).

Simultaneously, however, as the policy emphasis placed on outcome-based measures of teacher quality has increased, principals have had to justify their observational evaluations in the context of apparently objective, score-based rankings. For example, it is harder to defend high observational ratings of teacher quality when test score data suggest student performance is static or declining. Test-based measures, such as value-added models or growth-over-time models, ostensibly provide an external “objective” judgment of the effectiveness of teachers relative to other teachers in the district. Test-based rankings may confirm ratings assigned by principals, but they may also call them into question. For example, an “objective” measure potentially diminishes the authority of principals, if and when a principal’s ratings do not seem consistent with
what one might expect based on test-based measures. Evidence on this is scant, but Jacob and Lefgren (2008) found evidence that the test-based accountability environment in which principals and teachers operate may influence their assessments of teacher effectiveness. Specifically, Jacob and Lefgren (2008) noted that principal observations may be biased by two phenomena: (a) they may be more focused on low-performing, as opposed to high-performing, students; and (b) they may be focused on achievement levels (as defined, perhaps, by NCLB) as opposed to how much value a teacher adds to the students’ learning. In other words, they are more focused on the percent proficient than on the extent of gains, which may reflect the policy bias toward emphasis on proficiency thresholds.¹ This suggests that how principals operationalize their assessments of teacher quality may in part be influenced by how test-based accountability policies define teacher quality.

Current Study

This study is a cross-case analysis of how principals in one large urban district, LAUSD, use and understand multiple measures of teacher quality in the teacher evaluation process (see Appendix B). As Campbell (1984) observed, “There is the mistaken belief that quantitative measures replace qualitative knowledge. . . . Without competence at the qualitative level, one’s computer printout is misleading or useless.” I chose to pursue a qualitative study of principal evaluation of teachers in the context of a multi-measure teacher evaluation plan, because very little is known about how principals actually use and understand these tools in their practice. In contrast to international

¹ At the time of the study, the teachers union was petitioning for a proficiency approach, rather than a value-added approach, for use in teacher evaluations.
approaches to assessing development of teachers, research in the United States has focused on experimental studies of impact with little attention to the “contexts and processes by which such impacts are achieved” (Thomas & Pring, 2004, p. 56). Thus, my goal in this exploratory study is to provide a fine-grained descriptive analysis, based on principal interviews, to increase our understanding of principals’ behaviors and the mechanisms and processes of how they understand and use teacher evaluation tools in LAUSD. This exploratory research can inform subsequent causal research on the effects of multi-measure teacher evaluation models.

Specifically, this study examines how principals in LAUSD’s pilot Educator Growth and Development Cycle understood and used both a test-based measure and a standards-based observation protocol to evaluate teachers and lead improvements in instructional practice. In this pilot, the LAUSD invested heavily to develop a test-based measure—the Academic Growth over Time—that ranked teachers against other teachers with similar students within the district, based on the test performance of their students. In addition, the District trained and supported all participating principals in the use of the Teaching and Learning Framework, a standards-based protocol for evaluating teacher practice. These principals were positioned to analyze and understand the relationship between multiple measures of teacher effectiveness and the policy goal of improving teacher quality, and by extension, student learning. This study illuminates these perspectives and identifies potential implications for how principals operationalize use of growth measures and observation-based rating tools in practice, and provides insight into whether and how they use these tools to improve instruction across a variety of contexts.
In this study, I found that for the most part, principals expressed substantial concerns about whether the test-based measure reliably and accurately captured the true “value” that all teachers brought to the learning of their students. This skepticism made principals cautious about relying on or using the AGT rankings in any significant or consequential way, unless mediated by their professional judgment. The exception was a fairly young principal in a high-poverty school, who used the AGT as a tool for holding teachers “accountable” in the face of high teacher turnover and low student scores.

In contrast, principals universally described the TLF as a powerful tool for capturing evidence of teacher effectiveness and providing teachers with specific, actionable feedback they could use to get better. While principals were clearly ambivalent and sometimes dismissive of the usefulness of the AGT, they were positive about the TLF, and in many cases, worked hard to implement it with fidelity in the face of enormous challenges.

While principals felt the TLF provided teachers with extraordinary support and guidance as they worked to become more effective, it was also clear that the time demands and focus needed to implement at scale were hard to come by in some of the highest-poverty schools in the sample. In these schools, student needs and emergencies pulled principals away from a focus on instruction with striking frequency, even in the course of the interviews. This study thus also suggests that good policy and practices are not enough to improve learning. The demands on the principals in the highest-poverty schools in this sample suggest that socioeconomic segregation and associated disadvantages may undermine strong and thoughtful efforts to improve instruction.

I begin in Chapter 2 with a discussion of the study site and sample. This includes
an overview of LAUSD’s Educator Growth and Development Pilot, in which principals were trained to use a test-based growth measure (AGT) as well as the TLF to evaluate teachers and support improvements in instruction. Chapter 3 details study methods. Chapter 4 provides a discussion of what principals thought LAUSD’s AGT rankings measured, and additionally, how accurate and consistent they thought the AGT was as a measure of relative teacher quality. Chapter 5 explores how principals reported using the AGT in their support of teachers, and how this use was mediated by their understanding of the AGT. In Chapters 6 and 7, I discuss the principals’ perceptions of the reliability and accuracy of teacher ratings based on the TLF, as well as the role they felt the TLF played in improving instruction. Chapter 8 explores how principals characterized the relationship between the AGT and the TLF, and in particular, how they made sense of discrepant cases, in which they found the AGT ranking surprising, given their knowledge of a teacher’s instructional practice. I conclude with a discussion of the policy implications of this work.
Chapter 2: Context

Los Angeles and the Educator Growth and Development Program

In the 2011–12 school year, LAUSD launched a pilot for a new process for professional supervision, called the Educator Growth and Development Cycle (EGDC), which incorporates diverse measures of teacher quality in the evaluation process. In 2012–13, the district introduced a growth measure (the Academic Growth over Time rankings) and a standards-based educator observation tool (the Teaching and Learning Framework), and invited administrators to complete a weeklong training and certification process related to the TLF. The purpose of the pilot was to test and tune tools and practices, with the goal of full implementation as the evaluation system of record for accountability purposes in the 2013–14 school year.

Academic Growth over Time Rankings

The Academic Growth over Time measure is a statistical measure that is intended to estimate the impact of an individual school, principal, or teacher on student learning, as measured by the California State Tests (CSTs). CSTs purport to “measure students' progress toward achieving California's state-adopted academic content standards in English–language arts (ELA), mathematics, science, and history–social science, which describe what students should know and be able to do in each grade and subject tested.” For this study, I interviewed principals about teacher-level AGTs, which are a measure of the impact of individual teachers on their students’ learning over most of a year of instruction. The AGT is a form of value-added model that first predicts a student’s CST
score in a given year, based on his or her score in a prior year (or years, if available), as well as a set of his or her demographic characteristics. Then, to generate an AGT ranking, the model compares how well a teacher’s students score, compared to how well they were predicted to score.

**Academic Growth Over Time**

If the student performs higher than predicted, it is a positive value-added result.

If the student performs lower than predicted, it is a negative value-added result.

Figure 1: Logic of the AGT measure, as articulated in district materials

If a teacher’s students scored better than predicted, compared to similar students district-wide, then the teacher had a higher AGT ranking than other teachers with similar students who did not score better than predicted. If a teacher’s students scored worse than predicted, then the teacher’s AGT was lower than that of other teachers in the district whose similar students scored as predicted or better. Teachers were ranked in five performance categories based on how much their students’ actual scores differed from
predicted scores: Far Above Predicted, Above Predicted, Within the Predicted Range, Below Predicted and Far Below Predicted.

**The Teaching and Learning Framework**

The Teaching and Learning Framework is a standards-based framework for teaching that identifies key standards of effective practice and components within the standard. In a rubric form, the Framework provides descriptions of what practice looks like at four different levels of effectiveness for each component. For example, in Figure 2 below, the TLF suggests that a teacher who is “effective” at communicating the purpose of a lesson is a teacher whose “explanation of the instructional purpose is clear to students, including connections to big ideas and essential understandings. Most students are able to communicate the purpose of the lesson to their peers and others.” Principals script what they see and hear, and match this evidence to descriptions of practice to determine what level of performance a teacher has demonstrated.

---

2 The TLF is a modified version of the Charlotte Danielson Framework for Teaching, which is described on the Danielson Group’s website as “a research-based set of components of instruction, aligned to the INTASC standards, and grounded in a constructivist view of learning and teaching.”
### Standard 3: Delivery of Instruction

**Component 3a: Communicating With Students**

The presentation of a lesson impacts its outcome. In order to successfully engage students in the lesson, teachers need to clearly frame the purpose of the lesson including presenting the context. Teachers must communicate reasonable and appropriate expectations for learning, provide directions and describe procedures with clarity, model and expect the use of academic language, and use multiple strategies to explain content to meet diverse student learning needs.

<table>
<thead>
<tr>
<th>Element</th>
<th>Ineffective</th>
<th>Developing</th>
<th>Effective</th>
<th>Highly Effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>3a1. Communicating the Purpose of the Lesson</td>
<td>Teacher does not explain the purpose of the lesson or the purpose is communicated as only a series of activities/directions.</td>
<td>Teacher attempts to explain the purpose of the lesson, however, the instructional purpose is not clear or connected to big ideas and essential understandings. Most students are unable to communicate the purpose of the learning.</td>
<td>Teacher’s explanation of the instructional purpose is clear to students, including connections to big ideas and essential understandings. Most students are able to communicate the purpose of the lesson to their peers and others.</td>
<td>Teacher’s explanation of the instructional purpose is clear to students. Students articulate connections to big ideas, essential understandings and/or real-world application. Students are able to communicate the purpose of the lesson to their peers and others.</td>
</tr>
</tbody>
</table>

Figure 2: Sample Standard, Component, and description of practice from LAUSD’s Teaching and Learning Framework

In the initial stages of the pilot, LAUSD received feedback that the full Framework was challenging and complex to administer at scale, given other demands on principals’ time. In response, LAUSD narrowed the standards in the operational framework to a set of high-priority elements (see Figure 3).
LAUSD TEACHING AND LEARNING FRAMEWORK 2014-2015 Focus Elements

<table>
<thead>
<tr>
<th>STANDARD 1: PLANNING AND PREPARATION</th>
<th>STANDARD 2: CLASSROOM ENVIRONMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a. Demonstrating Knowledge of Content and Pedagogy</strong></td>
<td><strong>a. Creating an Environment of Respect and Rapport</strong></td>
</tr>
<tr>
<td>1. Knowledge of Content and the Structure of the Discipline</td>
<td>1. Teacher Interaction with Students</td>
</tr>
<tr>
<td><strong>b. Demonstrating Knowledge of Students</strong></td>
<td><strong>b. Establishing a Culture for Learning</strong></td>
</tr>
<tr>
<td>1. Awareness of Students’ Skills, Knowledge, and Language Proficiency</td>
<td>1. Importance of the Content</td>
</tr>
<tr>
<td>2. Knowledge of How Children, Adolescents, and Adults Learn</td>
<td>2. Expectations for Learning and Achievement</td>
</tr>
<tr>
<td>3. Knowledge of Students’ Special Needs</td>
<td>3. Student Ownership of their Work</td>
</tr>
<tr>
<td>4. Knowledge of Students’ Interests and Cultural Heritage</td>
<td><strong>c. Managing Classroom Procedures</strong></td>
</tr>
<tr>
<td><strong>d. Designing Coherent Instruction</strong></td>
<td>1. Management of Routines, Procedures, and Transitions</td>
</tr>
<tr>
<td>1. Standards-Based Learning Activities</td>
<td>2. Management of Materials and Supplies</td>
</tr>
<tr>
<td>2. Instructional Materials, Technology, and Resources</td>
<td>3. Performance of Non-Instructional Duties</td>
</tr>
<tr>
<td>4. Lesson and Unit Structure</td>
<td><strong>d. Managing Student Behavior</strong></td>
</tr>
<tr>
<td><strong>e. Designing Student Assessment</strong></td>
<td>1. Expectations for Behavior</td>
</tr>
<tr>
<td>1. Aligns with Instructional Outcomes</td>
<td><strong>e. Monitoring and Responding to Student Behavior</strong></td>
</tr>
<tr>
<td>2. Planning Assessment Criteria</td>
<td>2. Monitoring and Responding to Student Behavior</td>
</tr>
<tr>
<td>3. Design of Formative Assessments</td>
<td><strong>STANDARD 3: DELIVERY OF INSTRUCTION</strong></td>
</tr>
<tr>
<td>4. Analysis and Use of Assessment Data for Planning</td>
<td><strong>a. Communicating with Students</strong></td>
</tr>
<tr>
<td><strong>STANDARD 5: PROFESSIONAL GROWTH</strong></td>
<td>1. Communicating the Purpose of the Lesson</td>
</tr>
<tr>
<td><strong>b. Participating in a Professional Community</strong></td>
<td>2. Directions and Procedures</td>
</tr>
<tr>
<td>1. Relationships with Colleagues</td>
<td>3. Delivery of Content</td>
</tr>
<tr>
<td>2. Promotes a Culture of Professional Inquiry and Collaboration*</td>
<td><strong>b. Use of Academic Language</strong></td>
</tr>
<tr>
<td><strong>c. Demonstrating Professionalism</strong></td>
<td><strong>c. Using Questioning and Discussion Techniques</strong></td>
</tr>
<tr>
<td>1. Ethical Conduct and Compliance with School, District, State, and Federal Regulations</td>
<td>1. Quality and Purpose of Questions</td>
</tr>
<tr>
<td>2. Advocacy/Intervention for Students</td>
<td>2. Discussion Techniques and Student Participation</td>
</tr>
<tr>
<td>3. Decision-Making</td>
<td><strong>d. Using Assessment in Instruction to Advance Student Learning</strong></td>
</tr>
</tbody>
</table>

Highlighted Elements are identified as the Focus Elements for the 2014-2015 School Year.

* Evidence of teacher practice for this element will be collected for the Contributions to School Community Pilot for the 2014-2015 School Year.

Figure 3: Focus elements of LAUSD’s Teaching and Learning Framework
Because the modified framework is not identical to the original Danielson Framework for Teaching (Danielson, 2007), it is possible that any positive effects associated with use of the Danielson Framework in prior research (Gallagher, 2004; Kimball et al., 2004; Milanowski, 2004; Milanowski, Kimball, & Odden, 2005) will be attenuated in Los Angeles.

To participate in the EGDC pilot and use the TLF for observations, principals had to participate in a weeklong training, including three calibration exercises. The purpose of this calibration was to ensure that principal ratings of teacher effectiveness using the TLF were reliable, unbiased, and accurate. High-quality calibration increased the likelihood that the rating a principal assigned to a teacher was the same rating a different principal with the same training and using the same framework would have assigned to that same teacher, and also that the rating was the same rating they would have assigned on a different occasion of rating for teaching of the same level of quality. Without this kind of consistency, the TLF ratings cannot be used to compare teacher effectiveness across schools or across time, nor can they be counted upon to accurately identify levels of teacher efficacy within schools.

**Reconciliation of the AGT and the TLF**

Internally, based on initial data, LAUSD found modest correlations of about 0.3 between TLF ratings and AGT rankings (Bookman, 2012). However, it is not yet clear what explained the low correlations or whether principals observed a relationship between the two measures in their own work. Given that the two measures ostensibly both measured teacher quality, one might question why the correlation was not stronger.
This present study asked principals to comment on what they expected the relationship to be between the AGT rankings and the TLF ratings.

**Accountability Context**

This study took place in the context of federal, state, and local accountability models, which all incorporated some elements of high-stakes, test-based accountability. Because this study examined principals’ understanding and use of a test-based measure, it is relevant to understand that the basis for this measure—student scores on the CSTs—was also used for consequential purposes beyond what is explored in the study. In other words, data from the same test were used to evaluate teachers, principals, and schools for local, state, and federal purposes. Thus the CST was at the center of high-stakes decisions made by a variety of policy entities, many of which shared a goal of demonstrating improvement. Various uses of the scores may have influenced or incentivized behaviors to improve scores, some more productive than others. Some of the test-based measures used in LAUSD are explained below.

**Proficiency Thresholds and “Percent of Students Scoring as Proficient” (Federal Accountability)**

Under the federal No Child Left Behind Act, schools were evaluated based on the percent of their students who scored as proficient on the CSTs. Proficiency thresholds in a given year or on a given subject test were set without reference to thresholds in other years or other subjects. Thus, the threshold falls in different places in the distribution of test takers on each test, so comparisons should not be made across grade levels and
subjects, even though they often are. Proficiency scores are summative measures of achievement, in that scores are supposed to “sum up” what students know and can do at a given point of time. The CST samples a subset of the domain outlined in the state standards for each grade level, and student performance on this sample is the basis for inferences about student mastery in the entire domain outlined in the sample. Because some standards are more difficult to assess in standardized tests than others (such as inquiry), the tests may predictably emphasize some standards and de-emphasize others.

At the time of the study, about 23% of LAUSD elementary schools met proficiency targets for all subgroups, while only 6% of middle schools met targets. Under federal guidelines, in Los Angeles as elsewhere, schools that received federal Title I funds and did not make AYP criteria for two consecutive years were subject to identification for Program Improvement (PI). The number of schools entering PI status in LAUSD increased in successive years, as it did in the year of the study. This created a strong incentive for school staff to focus on the percent of students who scored proficient in each of the federally designated subgroups.

At the time of the interviews, LAUSD was exploring applying for a “flexibility waiver” from the U.S. Department of Education in exchange for using test scores for teacher evaluation.

Academic Performance Index (State Accountability)

The state’s Academic Performance Index (API) is a composite score based on the combined results from a variety of sources, including the CSTs in grades 3–11, as well as from modified versions of these tests administered to students with disabilities. Results
from the California Exit Exams are incorporated in high school scores; however, the schools in this study did not include any high schools. API scores range from 200 to 1000, and the statewide target for all schools is 800 for all students, and for all subgroups of students within each school. In Los Angeles in 2012, a quarter of schools scored between 750 and 799, and 37% scored 800 or above. Only 9% of schools (58) scored below 650, including three in this study’s sample.

In addition to rating schools based on the overall performance of students, the state’s model evaluated schools based on their API growth. Specifically, each school and each subgroup within each school is expected to show growth each year toward the state’s API target of 800. Each school’s growth target for a given year is 5% of the gap between the school or subgroup’s API score in the preceding year and the district target of 800. Note that the growth target will necessarily be larger in each year for schools that are farther from the district target; thus, to meet state growth guidelines, low-performing schools and subgroups need to make larger gains than higher-performing schools. Low-performing schools and subgroups therefore might be expected to face greater pressures to demonstrate improvements, and might be expected to place more emphasis on improving scores.

School Performance Framework (LAUSD Accountability)

LAUSD developed a School Performance Framework (SPF) to evaluate and provide data on school performance with respect to student achievement in terms of both status (e.g., level of performance) and growth (e.g., gains in performance). The espoused goals of the SPF were:
• to support and improve overall student learning and achievement;
• to provide the most complete and comprehensive picture of how schools in LAUSD are performing; and
• to assist teachers and school site leaders in setting school goals that are aligned with the Single Plan for Student Achievement (SMART Goals) (LAUSD School Performance Framework Guide, 2003).³

Germene to this study is the fact that both the status and the growth measures incorporated in the SPF were based on performance on a single set of exams: the CSTs. Student performance was defined in terms of performance levels on the CST, while growth points were awarded based on the cohort-to-cohort change in scores from one year to the next in the same grade and subject.

³ http://schoolinfosheet.lausd.net/budgetreports/schperfreports.jsp
Chapter 3: Research Methods

This study is a cross-case analysis of how principals in one large urban district, LAUSD, use and understand multiple measures of teacher quality in the teacher evaluation process. I chose to complete a qualitative study to explore:

1. What do principals think the AGT and the TLF measure, and what do they think is the relationship between these tools?
2. How does this understanding shape their implementation and use of these tools, both to evaluate teacher quality and to improve instructional practice?
3. Do these tools change the role of principals, and if so, how?
4. Does the context in which a principal operates (e.g., a high- or low-scoring school) shape which tools he or she uses or values?

In this section I discuss the study site, sample, and research design.

Sample

LAUSD provided me with the names and contact information of 150 principals in elementary and middle schools from the pilot phase of their own implementation of the above-described teacher quality initiative. These principals were volunteers. The benefit of participating in the pilot was having a full year to use the tools in a lower-stakes way before they were used district-wide. From the list, I selected an initial 25 elementary or middle school principals for potential interviews. I selected principals from elementary and middle schools, because initial data from the pilot suggested that implementation at the high school level is more problematic and less complete. For example, the use of
growth measures at the high school level has been confounded in part by the challenge of isolating the effect of an individual teacher from other teachers who teach the same student in other subjects. The elementary schools in the pilot represent a “best case” scenario for fidelity of implementation.

To ensure a somewhat diverse sample, I selected these principals by drawing a grid on a map of Los Angeles, and then selecting schools from different quadrants. This strategy capitalized on the relatively high degree of racial and socioeconomic segregation in the city, and resulted in a proposed sample that included principals from schools with a diverse set of demographics and operating structures and current levels of performance (see Figures 4 and 5).

Potential interviewees were first contacted by email, and then by follow-up call. I interviewed those who agreed to participate. I interviewed principals during two separate trips to Los Angeles, the first right before the start of the school year, and the second in the winter. During the first wave some interviewees were in training or otherwise unavailable. In both waves, some were nonrespondents, and some responded after the interview window. Some responded initially, then never confirmed an appointment time before the end of the trip. In the second wave, which occurred in winter, one principal had to cancel an appointment on the day of the interview due to a health emergency and another missed the interview due to a family death. Principals were more likely to decline in the second wave, which occurred while school was in session. As invitees declined to participate or were nonrespondents, I selected additional invitees following the same purposive strategy. Out of 38 invitees, I was able to interview 17 interviewees. Principals
in both waves were tightly scheduled, but principals in the highest-performing schools had the greatest scheduling flexibility.

I sampled purposively to secure a sample that represented principals in schools that were diverse both in terms of demographics and in terms of performance, because demographics can influence school performance, and school performance can influence the responses of teachers and principals to accountability measures. For example, students who frequently transfer among schools during the school year are at greater risk of academic or behavioral challenges (Schwartz, Steifel, & Cordes, 2015).

Because interviews were conducted during two separate trips to Los Angeles, I also interviewed two principals twice, once on each trip, to see if policy change in between waves substantively changed the content of their interviews and to use the second interview to play back and test themes and patterns that emerged in the first round of interviews. I offered interviewees gift cards worth $50 for their participation in the interview. Some initially declined the cards. Some of those later accepted the cards when told they could use them on items for their schools or staff.

The principals in the sample represented schools that were diverse in demographics and performance. For example, the percent of students coming and leaving within the school year ranged from 1% in one of the highest-performing schools to almost 50% in one of the lowest-performing schools (see Figure 4). Similarly, the percent of students living in poverty for each school ranged from 23% to 100%. Most schools in the sample served predominantly Latino students, or in one case, African American students. However, three sites had a more heterogeneous composition. High levels of student turnover affect the ability of teachers to establish and maintain classroom culture
and sustain instructional momentum in ways that are different than in classrooms where student turnover is low. Factors like turnover and adverse poverty potentially affect the ability of a teacher to improve student learning in ways that might affect the estimates of teacher efficacy. Those factors are represented in this sample.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>% Low income</th>
<th>% Latino</th>
<th>% Black</th>
<th>% ELL</th>
<th>% Entering and leaving</th>
<th>Accountability Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr. Isaac</td>
<td>ES</td>
<td>55</td>
<td>51</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>Excelling</td>
</tr>
<tr>
<td>AVG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ms. Kim</td>
<td>ES</td>
<td>23</td>
<td>17</td>
<td>11</td>
<td>9</td>
<td>11</td>
<td>Achieving</td>
</tr>
<tr>
<td>Ms. Alana</td>
<td>ES</td>
<td>23</td>
<td>19</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>Achieving</td>
</tr>
<tr>
<td>AVG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>23</td>
<td>3 to 11</td>
<td></td>
<td></td>
<td></td>
<td>from 1 to 11</td>
<td></td>
</tr>
<tr>
<td>Mr. Roland</td>
<td>ES</td>
<td>10</td>
<td>100</td>
<td>0</td>
<td>38</td>
<td>18</td>
<td>Service and support</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Service and support</td>
</tr>
<tr>
<td>Ms. Suss</td>
<td>ES</td>
<td>80</td>
<td>89</td>
<td>2</td>
<td>24</td>
<td>10</td>
<td>Service and support</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Service and support</td>
</tr>
<tr>
<td>Ms. Charles</td>
<td>ES</td>
<td>10</td>
<td>80</td>
<td>6</td>
<td>25</td>
<td>12</td>
<td>Service and support</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Service and support</td>
</tr>
<tr>
<td>Ms. Dale</td>
<td>MS</td>
<td>69</td>
<td>53</td>
<td>5</td>
<td>11</td>
<td>12</td>
<td>Service and support</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Service and support</td>
</tr>
<tr>
<td>Mr. Frank</td>
<td>ES</td>
<td>79</td>
<td>96</td>
<td>0</td>
<td>52</td>
<td>20</td>
<td>Service and support</td>
</tr>
<tr>
<td>AVG</td>
<td>69</td>
<td>0 to 6</td>
<td></td>
<td></td>
<td></td>
<td>from 10-20</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>69</td>
<td>0 to 6</td>
<td></td>
<td></td>
<td></td>
<td>from 10-20</td>
<td></td>
</tr>
<tr>
<td>Ms. Rosa</td>
<td>ES</td>
<td>10</td>
<td>97</td>
<td>0</td>
<td>43</td>
<td>13</td>
<td>Watch</td>
</tr>
<tr>
<td>Ms. Richmond</td>
<td>MS</td>
<td>76</td>
<td>80</td>
<td>6</td>
<td>15</td>
<td>10</td>
<td>Watch</td>
</tr>
<tr>
<td>Ms. Ivy</td>
<td>ES</td>
<td>10</td>
<td>95</td>
<td>2</td>
<td>64</td>
<td>23</td>
<td>Watch</td>
</tr>
<tr>
<td>Mr. Thom</td>
<td>MS</td>
<td>10</td>
<td>99</td>
<td>0</td>
<td>17</td>
<td>16</td>
<td>Watch</td>
</tr>
<tr>
<td>Ms. Brown</td>
<td>ES</td>
<td>10</td>
<td>90</td>
<td>3</td>
<td>38</td>
<td>23</td>
<td>Watch</td>
</tr>
<tr>
<td>Ms. Irma</td>
<td>MS</td>
<td>84</td>
<td>98</td>
<td>1</td>
<td>20</td>
<td>19</td>
<td>Watch</td>
</tr>
</tbody>
</table>
The principals in the sample led schools that appeared diverse with respect to accountability and performance as well (see Figure 5). Because the empirical literature on high-stakes testing suggests that high-poverty and lower-scoring schools may respond differently to test-based accountability than higher-scoring or more affluent schools, I wanted to make sure both types of environments were included in the sample.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>% Proficient ELA</th>
<th>% Proficient math</th>
<th>API SY12</th>
<th>API change</th>
<th>Growth points</th>
<th>Accountability Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr. Isaac</td>
<td>ES</td>
<td>80</td>
<td>91</td>
<td>931</td>
<td>17</td>
<td>3</td>
<td>Excelling</td>
</tr>
<tr>
<td>AVG</td>
<td></td>
<td></td>
<td></td>
<td>931</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Ms. Kim
- **Achieving**

| Service and support | 47 | 4 | Service and support | 31 | 2 | Service and support | 2 | 3 | Service and support | 17 | 3 |
|---------------------|----|---|---------------------|----|---|---------------------|----|---|---------------------|----|---|---------------------|----|---|---------------------|----|---|---------------------|----|---|
| **Range**           | 942|   | **Range**           |   |   | **Range**           |   |   | **Range**           |   |   |

### Ms. Alana
- **Achieving**

### AVG

### Ms. Rosa
- **Watch**

### Ms. Richmonds
- **Watch**

### Ms. Ivy
- **Watch**

### Mr. Thom
- **Watch**

### Ms. Brown
- **Watch**

### Ms. Irma
- **Watch**

### Mr. Grace
- **Watch**

### AVG

### Range
To check how the performance of the schools of the principals in the sample compared to the performance of schools in LAUSD as a whole, I compared the performance of sample schools to schools in LAUSD using the state’s API rankings. The number of schools in the sample is small, so the percentages might vary substantially with slight changes in composition of the study sample; but overall, the performance composition of the sample seems to roughly parallel that of LAUSD as a whole (see Figure 6).

<table>
<thead>
<tr>
<th>% Scoring</th>
<th>LAUSD</th>
<th>Study Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>800 or above</td>
<td>37%</td>
<td>24%</td>
</tr>
<tr>
<td>Between 750 and 799</td>
<td>25%</td>
<td>24%</td>
</tr>
<tr>
<td>650-749</td>
<td>29%</td>
<td>30%</td>
</tr>
<tr>
<td>Below 650</td>
<td>9%</td>
<td>18%</td>
</tr>
</tbody>
</table>

Figure 6: Distribution of schools scoring at different API levels in LAUSD and the study sample (data excludes ASAM, charter, and small schools)

**Data**

In this study, I rely on two kinds of data: documentation and data gathered from interviews.
**Documentation.** I gathered and reviewed publicly available documentation on:

- the Educator Growth and Development Cycle;
- the Academic Growth over Time measure;
- the Teaching and Learning Framework (LAUSD’s modified version of the Danielson Framework for Teaching);
- school report cards for the schools where interviewees were principals;
- news stories related to the EGDP model;
- news stories about the lawsuit related to the use of test data for teacher evaluation; and
- news stories related to the LA Times value-added model, which preceded the AGT and has been controversial, and which in fact was a topic of discussion in several interviews.

**Focused interviews.** Interviews allowed me to focus data gathering on the target of inference and give me direct access to principals’ self-reported understanding of what each of the teacher efficacy measures (TLF and AGT) actually measured, as well as the relationship between these process and outcome measures of teacher efficacy (e.g., the observations using the TLF ratings and the AGT rankings) (see “Question Guide” in Appendix C). I constructed the interview protocols by researching the measures, and engaging in informal conversations with contacts in California schools, as well as administrators within the district central office.

Questions were designed to test and explore several propositions regarding how principals understood the teacher evaluation tools they used, as well as the relationship between this understanding and how they chose to use these tools to improve instruction.
The existing knowledge base on principal use of multiple measures for teacher evaluation is scant, but enough is known about principal observations and test-based accountability to hypothesize that the effect of AGT data on principal understanding of teacher quality will be influenced or mediated by principal understanding of measurement principles and familiarity with both the form and substance of the tests that are the basis for AGT scores. In addition, the success of implementation depends both on the value those responsible for implementation place on the policy they must implement, as well as on the resources and capability they have to implement the policy with fidelity. Interview questions explicitly addressed what principals thought the AGT and the TLF measured, as well as how consistent and accurate they thought these tools were as measures. Questions also explored how principals felt these tools related to their purpose and work as leaders of instructional improvement.

In the first wave of interviews, I began with several hypotheses about how principals understood and used the AGT rankings and the TLF in the process of teacher evaluation (see Appendix A). I assessed this set of propositions—what Sutton and Straw (1995) would call hypothetical stories “about why acts, events, structure and thoughts occur” (p. 378)—against what principals actually say about their experience using multiple measures to assess teacher quality. In the next wave of interviews, I probed principals during interviews to test how they made sense of or interpreted the AGT rankings in the context of their perceptions of teacher quality, based on their observations using the TLF. In the conclusion of this study, I discuss the implications of these findings for the validity of inferences about teacher quality based on TLF ratings and AGT rankings.
Interviews occurred at the participants’ convenience in Los Angeles, California, during August 2012 and January–February 2013. Each interview lasted between 45 minutes and two hours, with most lasting about an hour and a half.

Data Collection and Audit Trail

I collected data in two waves, using a semi-structured interview approach. Miles and Huberman (1994) noted that the purpose of field research is “to describe and analyze a pattern of relationships” (p. 17). After each day of interviews, I wrote field notes in which I summarized emerging themes and questions, including patterns in principals’ descriptions of the AGT and the TLF, as well as opportunities and challenges they saw with respect to the uses of these tools.

I had as a primary purpose understanding how principals situate themselves with respect to these measures and their use, and how this in turn mediates implementation. To this end, as I analyzed preliminary interview data, after the first round of interviews, I coded emerging themes and then checked and validated these themes by asking subsequent interviewees about the relevance and explanatory power of these themes (Cooney, 2011; Strauss & Corbin, 1998). I also used these data to tune my interview protocol—to add or modify questions. Because my focus is on the behaviors or understandings of my subjects, there is little purpose in trying to corroborate what they say with other sources of information. However, to ensure accuracy, I recorded all interviews and offered interviewees the opportunity to review and append transcripts. In reporting data, as requested by principals, I only made minor changes to fix minor
grammatical conventions that are appropriate in casual conversation but appear awkward in written and more formal contexts.

Data Analysis

My analytic strategy was to pursue a cross-case analysis (Yin, 2009), with each principal treated as a separate case. I approached analysis of the data iteratively. First, I read each transcript carefully, and checked or verified sections where the transcript seemed unclear. Then, I used a qualitative data analysis software program (HyperresearchTM) to facilitate coding and analysis. To support this process, I developed an initial list of etic codes, based on the literature and my initial propositions about how principals understand and use the measurement tools (the AGT and the TLF) (see Appendix E). Based on this preliminary review, as well as conversations with key stakeholders (including leaders of EGDP and the Department of Talent Management in LAUSD), I revised my protocol and amended my code list with any emic codes, or codes that emerged from the data in the preliminary review process. I then conducted a second round of interviews of ten principals in LAUSD.

After interviews were complete, transcribed, and coded in a preliminary way, I used a strategy of pattern matching (Yin, 2009) to analyze the fit of principals’ responses to my initial propositions about principal use of multiple measures of teacher efficacy. Miles and Huberman (1994) recommended manipulating the data for analysis by organizing it in different ways. I contrasted interviewees by whether or not they served in high- or low-poverty schools (which also tended to be high- and low-performing schools, respectively), and whether principals in these categories differed in their understanding of
the AGT and the TLF, in their reported comfort with these tools, and in their reported use of these tools.

Corbin and Strauss (2008) recommended looking for “negative cases.” Similarly, Maxwell (2005) suggested examining “discrepant data.” My sample is small enough that I treated all differences as potentially substantive differences, rather than discrepant cases. In the context of this study, however, I asked principals about surprising findings, or incidents in which their assessment of quality using the TLF differed from what they would expect based on the AGT rankings, and then contrasted how principals in high- and low-performing contexts made sense of these discrepancies.

I also selected and presented interview excerpts that exemplify themes. In the second round of interviews, I followed up with two interviewees from the first round (Mr. Isaac and Mr. Roland), and played back themes that emerged in the initial interviews to ensure that my understanding of themes represented their experience.

Validity

This study is exploratory, in that it is designed to capture an exploratory cross-section of principal understandings and uses of the AGT rankings and the TLF ratings. In this, its goals are somewhat modest. It does not purport to be representative or comprehensive. It does attempt to capture a cross section of how different principals think. Thus, the primary validity burden is making sure the words and thinking of the principals is accurately captured and reflected, and not that the words and thinking are an accurate representation of all principals in the district. What the study can do is provide a portrait of how a carefully selected sample of principals made sense of and used the AGT
and the TLF. Any claim of causality or generalizability is beyond the scope of this study; that burden falls to future studies.

In this study, I do seek to address concerns with respect to four kinds of validity: construct validity, internal validity, external validity, and reliability (see Appendix G). In this study, as in all studies, the goal is to have the study findings be an accurate representation of reality (Maxwell, 2005).

I address construct validity primarily by ensuring that constructs that are the focus of interviews are developed from the literature. A potential threat to validity is that, if my questions were poorly framed or if interviewees simply stated what they perceived I wanted to hear, my findings would be inaccurate or biased (Yin, 2009). To address this threat, I used a semi-structured interview process (see “Protocol” in Appendix D) that allowed me to pursue my line of inquiry while posing questions in an open-ended, conversational way that enabled interviewees to explain how they made sense (Becker, 1998) of the AGT and the TLF, rather than challenging them to respond to my sense-making of those measures. During interviews, I frequently rephrased what I thought interviewees had said, so that they could confirm, revise, or build on my developing understanding of their ideas (Maxwell, 2005). I sent an interim summary of themes to my contacts in district leadership to check my set of emerging themes against their experience, and I also had several conversations with practitioners who were familiar with the district to verify that the constructs as I defined them reflected what they experienced in the field. These “member checks” with interviewees and practitioners ensured that although not all principals might agree with all opinions, the opinions expressed were consistent with what some principals might think.
As I analyzed data, I sought to ensure internal validity by engaging in pattern matching (Yin, 2009)—looking for patterns that emerge from the data. I first reviewed each case (a principal) for patterns that emerged from each case. I then compared across cases, to see if simpler findings were apparent across cases (external validity). I also matched my findings against my initial hypotheses, as well as alternative hypotheses about principal use and understanding of the measures. I then organized my data around themes, across cases. Finally, to minimize errors and biases, in such a way that my findings tell the same story that might be found by a second researcher interviewing the same subjects (reliability), I employed an interview protocol, wrote research memos summarizing key themes after each day of interviewing, and developed a case study database, which included transcripts of interviews, case study narrative notes written after each day of interviewing, and etic and emic codes used in analysis of data.

Finally, I shared my developing work extensively with colleagues, including practitioners and academics who have worked in related contexts or on related issues. The final work reflects their feedback.

To ensure that interviewees felt free to share thoughts openly, I assured them that their responses would be confidential. In communicating with me, they were invited to use private emails, and not their work contacts. I use pseudonyms for all interviewees throughout. It is possible that concerns about confidentiality might have affected the validity of responses, although I do not have any evidence to suggest or confirm this was the case. In fact, although the principals all had the opportunity to use their private email for all communications, most used school email accounts and expressed no concerns about being contacted at their schools.
Chapter 4: What Did Principals Think the AGT Ranking Actually Measures?

The AGT as a Measure of Teacher Quality

For principals to use the AGT rankings effectively for the purposes of teacher evaluation or teacher improvement, they must first be confident that the AGT measures important outcomes. Second, they must have confidence that the rankings are accurate and reliable and substantively meaningful. Overall, most principals in the sample appreciated the intent of the AGT, but were concerned that some student characteristics or conditions beyond what are captured in the AGT might affect scores in ways that distort the true underlying story about teacher effectiveness, and thus lead to misidentification of teachers as effective or ineffective. For this reason, most principals advocated that AGT rankings not be used in isolation from close contextual knowledge of the school environment and principals’ professional judgment.

Prior to the development of the AGT, the district evaluated school quality based on the overall performance, or status, of their students as measured on the CSTs. Schools that served higher-wealth students tended to fare better as higher-wealth students tend to score better, and inversely, schools with higher levels of poverty and student turnover scored lower and were ranked lower. Principals uniformly thought that status measures alone put a larger burden on schools with larger numbers of initially low-scoring students, because even if teachers in these schools substantially improved student learning, these schools still would have relatively low mean scores.

In contrast, principals valued the intent of the AGT, because it promised to measure the impact of schools and teachers on the learning of children, regardless of
where those students started academically and what advantages or disadvantages they had in life. Principals were aware that high-scoring students were not evenly distributed across schools, and as Ms. Dale noted:

There are schools that have a higher number of high-achieving students. They recruit them. It’s the area they live in. They’ve got parents that can have tutors. And so when you look at data, they always look good, but that doesn’t mean they’re moving the students.

Using the AGT, teachers who taught low-performing students could still receive a high ranking if their students gained more than similar students taught by other teachers. For this reason, most principals felt the AGT was an improvement on previous test-based metrics of teachers.

Principals in lower-scoring schools thought the AGT measure at least acknowledged that there were different challenges associated with teaching different students, and that the growth models help to distinguish their efforts from the corrosive effects of, in particular, poverty. As Mr. Grace explained: “The teachers want to believe that people understand what they deal with and the direction and the circumstances under which they are working.” The AGT gave principals in low-scoring contexts, like Mr. Grace, a way to keep a focus on student learning while acknowledging that the challenge of teaching and getting all students to a given level of performance was much greater in some contexts than others. They explained that the AGT could be used to hold schools and teachers accountable for improvement in scores, if it adequately accounted for differences in the skills, assets, and supports different students bring to the classroom.
Principals were less confident the AGT reliably and accurately provided substantively meaningful information about teacher efficacy—about how much of a difference individual teachers make in the learning of their students. Most principals, and particularly principals in higher-scoring schools or more affluent schools, identified at least one factor that might bias or erode the validity of inferences about teacher quality based on these scores. Principals who expressed a lack of confidence in the AGT as a measure of teacher quality were less likely to feel it was as important as other factors in their supervision of teachers. In contrast, principals in lower-scoring and higher-poverty schools were more likely than peers in higher-scoring or more affluent contexts to express confidence in the rankings as measures of the quality of the teachers they supervised.

To support high-quality supervision and evaluation, any measure, including the AGT, needs to be accurate and reliable. Several principals indicated they needed confirmation the AGT accurately identifies teachers who are effective or ineffective, before they could justify relying on it for evaluative purposes. Otherwise, they feared, any sanctions or feedback applied based on the AGT might be misdirected and might undermine efforts to achieve better outcomes for children.

Principals in the sample identified five conditions that might bias the AGT as a measure of the effectiveness of teachers in schools. First, many principals described purposeful placement practices, and pointed out that because these assignments are not random, they might systematically bias AGT rankings. Second, numerous principals identified student characteristics for which they felt the AGT statistical model might not adequately account. Third, several principals explained that they were not convinced the
model could separate and account for the impact of individual teachers, as distinct from effects associated with the larger climate of the school or class composition. For example, some noted that in more affluent schools, peers reinforce peers, whereas in high-poverty and high-turnover environments, transitions and deprivation of peers actually might limit learning. Fourth, several principals, all of them in high-scoring schools, described ceiling effects associated with the fact that a high proportion of their test takers earned perfect scores in successive years on the CSTs. It is reasonable to suspect that similarly, some schools that serve many English Language Learners or students growing up in very adverse circumstances might plausibly demonstrate floor effects, although nobody in the lowest-performing schools identified this as an issue. Finally, several principals described test-focused instruction that narrowly targets tested content and formats, at the expense of less tested or untested content or formats. This kind of teaching might lead to higher scores in the tested grades; however, these gains might not indicate mastery of more lightly tested or unsampled content in the standards. None of these factors was described by all principals, but every principal cited at least one of these factors.

In the next section, I examine how well principals in the sample think the AGT accurately and reliably measures what it purports to measure. Specifically, I analyze interviews to identify the evidence that principals provided related to potential bias, reliability, or validity of these tools.

Factors Principals Believe Shape the Accuracy and Reliability of the AGT Rankings

Class composition effects: Nonrandom assignment of students to teachers. In a 2008 study in the district from which this study’s sample was drawn, Kane and Staiger
(2008) compared teacher value-added estimates in schools to estimates from an experimental sample in which students were randomly assigned to teachers. The Kane and Staiger sample was unlike the sample in this study, in that almost all principals in this study sample described placement processes that led to nonrandom assignment of students to classrooms. In most cases, this involved teachers collaborating to ensure that students were placed in the following year with the teachers who were most likely to help them succeed. Principals stated that some of the most effective teachers might systematically teach the most challenging students, and this might affect their AGT ranking, though some were unclear on how. There were also noticeable differences in how principals in high- and low-scoring schools described placement practices.

In general, placement processes involve careful consideration of individuals and their specific needs, then assignment of students to teachers best qualified to address their needs, sometimes in clusters based on educational needs. Mr. Isaac, like several other principals, said he systematically placed students who were more difficult to “move” with teachers he perceived as his “best” teachers. These principals realized that the effect of these purposeful placement decisions might be to skew the AGT rankings of these teachers downward. Inversely, some of these principals speculated, teachers with “easier” loads might appear to be more “effective” than they really were. Several principals in the sample pursued this practice despite realizing that it might affect AGT rankings, because they felt it was in the educational interest of the children to do so. In some cases, these principals noted that this caused them to view the AGT rankings of some teachers with more skepticism, and to lean on their own professional judgment and observation in drawing inferences about teacher quality. Typically, they wanted the teachers with the
highest level of engagement and the greatest commitment to improving students’ learning to work with their most at-risk and vulnerable students.

For example, principals sometimes asked teachers they perceived as very strong to take challenging students in spite of the risk that doing so would adversely affect their AGT ranking. Mr. Isaac said he was able to make this request of his teachers because he was able to support them in other ways and because he was able to appeal to their professionalism. As he explained:

*We try to seek out those teachers that we feel are the best instructors to work with the at-risk students and go to them and say, “Will you take this on for us? This is important to our school.” We then support them, so that if they take [this responsibility] on, they get extra classroom aide time.*

Significantly, like other principals in higher-scoring schools, Mr. Isaac mentioned that his parents and community were not very focused on using the AGT to evaluate teacher efficacy. It is possible that principals in higher-scoring schools are able to systematically place students in ways that are better for children, even if they bias some teacher AGT rankings downward, precisely because their parent communities place less emphasis on the AGT as a measure of the effectiveness of their children’s teachers.

In her high-scoring school, Ms. Alana worked with her teachers internally to place students systematically with the teachers who they felt were most likely to be successful with each child, and kept these conversations a secret from parents, precisely so parents would not try to influence these decisions. As she explained:

*Our teachers know our students, and we feel like they know their fellow teachers, and they make a good match for them. So, for example, you might have a teacher...*
who’s really great with the highly gifted kids, just knows how to differentiate for them, right? So that teacher tends to get a certain type of student, and then that is reflected in her AGT.

In Ms. Alana’s high-scoring school, she said she trusted her teachers to ask for the students they felt would thrive in their classrooms, even though those requests might adversely affect their AGT ranking.

Like Mr. Isaac, Ms. Alana emphasized the critical importance of the principal’s professional judgment in drawing inferences about teacher effectiveness based on data. For example, Ms. Alana recalled a teacher in her building who demonstrated a unique ability to engage and support some of the most challenging students in the building. Ms. Alana described telling this teacher: “I can see why your scores have gone down. You’re getting the kids who are coming in with emotional whatever.” However, she acknowledged that somebody at Central Office or without that contextual knowledge might see those lower scores and come to a different conclusion about the effectiveness of that teacher. For that reason, she felt the AGT should not be used to make judgments about teacher quality in isolation from a principal’s informed professional judgment. As she explained: “If I don't have that in-house information, then I do think it’s sort of unfair to just bring in that below average prediction for this teacher, and say, ‘See, you’re not doing a good job.’”

Principals in some schools with overall scores that were average in the sample also described placing children with the teachers they thought would benefit them the most, but they seemed a little less confident about this practice. For example, like her colleagues in more affluent schools, Ms. Kim said her goal is to place “the more dynamic
teachers” with the students who have more difficulty learning, but she was acutely aware that this placed an extra burden on these teachers, particularly when the teacher ended up with a disproportionate share of very low-scoring or challenging students. She also mentioned that in her building currently, a couple of teachers she views as more “average” have lower performing and less engaged students who should not be in their classes, simply because they are not the teachers with the best skill and expertise to challenge and support those students.

While the high overall test scores in some settings seemed to insulate teachers from pressures associated with AGT rankings, principals in some of these lower or moderate scoring contexts acknowledged that some teachers in their buildings expressed reservations about accepting students who might suppress their AGT ranking. These principals felt that what was best for the child might not be best for the teacher, and struggled with the tension this created between the interests of teachers and the interests of children.

Not all principals with nonrandom placement practices were concerned about the effect on AGT scores of nonrandom placement. For example, Ms. Brown assumed that the AGT model would adequately capture the effect of those teachers, so long as they were able to support growth in their students. She justified her expectation based on her confidence in these teachers:

Some of it is just that these teachers have given these kids their self-confidence back, because they have that ability to do that. And that is a lot of it because some of these kids just have been deflated for one reason or another. So you’ve got to give them that self-confidence to know that they can move themselves.
In Ms. Brown’s view, these teachers might actually have stronger-than-expected AGT rankings compared to their peers, because they were uniquely skilled at connecting with hard-to-reach students and engaging them in learning.

Not all composition effects described by principals were the result of positive efforts to place challenging students with the teachers who would serve them best. Several principals described teachers who, they felt, sought to push challenging students out of their classes so that they would become the responsibility of other teachers. For example, Ms. Richmond described a teacher who was widely perceived by her administrative team as mediocre, but whose AGT rankings were consistently above average. In that case, Ms. Richmond questioned the accuracy of the AGT ranking, and her instinct as a principal was to go back and look closely at the students that teacher had taught in order to better understand the ratings the teacher received. In this case, for example, she examined course transfer data and concluded this teacher was not a teacher who ran “an academically rigorous classroom, but a teacher that’s very effective at weeding kids out.”

Examples like these led Ms. Richmond to intervene in course changes within her school. She noted that reducing the representation of unmotivated students in your classes potentially inflates your AGT ranking, and she cited this as the reason she now reviews every course change request. As she explained: “I began to look at why kids were being moved, and if it was because it was this teacher saying, ‘They’re lazy. They’re this. They’re that. They’re not doing homework.’” Ms. Richmond began to block placement requests when she felt the intent of the requests was to “cull” the class to a group of pliable and easy-to-teach students. However, if principals across schools are not equally
rigorous in monitoring course changes, this might erode the validity of inferences about teacher quality across school sites.

Ms. Island, the principal in the lowest-performing school in the sample, also was concerned that some teachers’ AGT rankings were a function of who they taught, not how they taught. She observed that when teachers shifted assignments and began teaching different kinds of students, their rankings based on test scores shifted. For example, she noted that two teachers had strong AGT ratings when teaching the school’s “gifted and talented” class, but when they were assigned lower performing groups, they lost the high rankings. Her inference was that somehow the AGT model did not adequately account for the effects of class composition, and that a ranking that shifts dramatically when class composition changes raises questions about whether “quality” can be measured independently of context and class composition.

Because of observations such as these, many principals wondered if the AGT might perversely reinforce or reward less effective teachers. To the extent the AGT incentivizes some teachers to influence student assignments in nonrandom ways, and to the extent that this nonrandom sorting differs across classrooms and sites, the value of the AGT ranking as a tool to compare teacher effectiveness is eroded.

Mr. Frank, a principal in a school with a high-poverty and high-ELL population, noted that when teachers shifted schools and taught different populations in a new context, their AGT rankings seemed to change. However, while Ms. Island questioned the AGT model, Mr. Frank wondered if the changes in rankings had more to do with teachers’ different effectiveness with different subgroups. Mr. Frank worked in a new school, and his teachers had, for the most part, moved there from teaching jobs in other
schools. He had access to data on AGT rankings for his teachers in both his current, high-poverty and high-ELL context and in their prior contexts. Mr. Frank described a teacher who received a very high AGT ranking after successfully teaching gifted and talented students in a preceding school, but who really struggled in the new school to be effective with English Language Learners, particularly in math, for which she was ranked as red (least effective) in the most recent year. Mr. Frank wondered if teaching less privileged children or teaching English Language Learners required a different skill or level of effort that is not necessary when teaching privileged children or native English speakers.

Mr. Frank felt he had to look at the teacher’s teaching to understand the AGT data and figure out aspects of her practice that might explain the change. He concluded this teacher needed to develop new strategies to better support a different kind of learner. Mr. Frank’s point was that “effectiveness” might be situational: some teachers may be very effective with some types of learners, and less effective with others. Mr. Frank’s observation is consistent with research by Loeb, Soland, and Fox (2014), which suggests some teachers are more effective with English Language Learners than other teachers.

Some principals also identified situations in which, in their opinion, parent preferences might also influence assignment of students across classrooms in ways that possibly biased AGT results. For example, Mr. Isaac also identified a set of teacher qualities and noncognitive outcomes that were valued by parents but not captured by the test, resulting in a dissonance between quality as measured by the AGT and quality as valued by his higher-performing school community. In turn, this perceived dissonance triggered efforts by parents to influence the placement of their students. For example, Mr. Isaac worried that the AGT might reward especially authoritarian and didactic teaching,
particularly in science. However, he also noted that authoritarian teaching was not valued by parents, and these parents often worked to avoid having their students in classes with such teachers. He explained:

Those teachers who are particularly harsh on the kids and are very demanding and not very patient and stuff, a lot of those teachers came back with very high value added scores . . . it just raises questions about whether the value added is really measuring something that at the end of the day we want to measure, because I can tell you, having placed students and dealt with parents, that 99 out of 100 of my parents would look at the value-added scores for those teachers and say, “Thanks, but no thanks. I’ll take my chances next door with the least effective teacher because I know that my child’s going to come to school and feel happy and say and it’s going to be a positive environment.”

Mr. Isaac explained that even with those highly effective ratings, his students’ parents have heard from other parents that those teachers are harsh and strict and work actively to not have their children placed in those classes. It was unclear how this systematic effort to avoid certain teachers might shape rankings, but this anecdote is also consistent with the possibility of nonrandom sorting of students across classrooms, in this case in response to parent preferences.

In sum, principals described both teachers’ actions and decisions and parental preferences that shape placement of students with teachers in nonrandom ways. Principals supported these preferences when they reflected a desire to place students with teachers who would engage and support them, or a desire to avoid placing them with teachers who—even if demanding—were rigid and inflexible and unconcerned with
personally engaging students. While some principals seemed to believe this latter, negative set of teacher qualities would be reflected in AGT rankings, many were not convinced. These skeptical principals felt strongly about the role of their judgment in interpreting the significance of AGT rankings.

Behind these concerns is the sense that in order to accurately identify teachers as effective or ineffective, the AGT model needs to accurately capture the complexity of the context and the particular students each teacher teaches, as well as the full breadth of good practice. With respect to whether AGT rankings are shaped by nonrandom sorting, many expressed a measured skepticism about the AGT and whether it alone could adequately account for the effects of nonrandom sorting, especially in the absence of the principal’s judgment and contextual knowledge.

**Inadequate modeling of relevant student characteristics.** Behind the discussion above about nonrandom sorting and the effect of class composition on teacher AGT rankings, is the implicit assumption that not all students are equally easy to teach and that teachers are more likely to generate strong academic progress within the year with some students rather than others. To the extent that the AGT model does not adequately account for all substantive differences that might affect student performance, its accuracy as a measure of teacher quality may be diminished. Principals in the sample cited several factors or student characteristics for which they felt the AGT model did not adequately account and which, in their opinion, made the AGT rankings in some cases less trustworthy than their own judgment. Overall, the interviews suggest that students in the lowest-scoring schools might face some unique challenges that adversely affect the likelihood their teachers will receive high AGT rankings.
One interesting perspective that came up in several schools is that the poverty measure in the AGT was too blunt and limited, and that this could have a substantial impact on how schools are ranked. Many principals in the sample, in one way or another, said that not all students in poverty are the same. For example, Mr. Isaac, a principal in a higher-scoring school, observed that even though on paper, 64% of his students were growing up in poverty, he felt the poverty in his school was qualitatively different than the concentrated, multi-generational poverty of some other schools. He compared his school (64% poverty) with other nearby schools with 70% or more of students living in poverty, and noted: “I think our students who live in poverty might have a lot more human capital versus the students in poverty in those other contexts, even though they’re characterized the same way by the metric.” In other words, even though students in his school came from families with limited financial means, he felt they had other resources or forms of social capital that might enhance performance in ways that were not true of families in poverty in some other schools. Because his school was a magnet, perhaps his parents had to demonstrate more initiative to enroll their children. Perhaps they were recent immigrants with intact families and invested parents, despite limited means.

Several principals cited the difference between stable family units with limited means and families that had been disrupted or broken apart, and mentioned that schools that serve families that are broken or have lived in poverty for multiple generations have unique challenges that make them substantively different from schools that may look similar in terms of demographic indicators.

Ms. Dale, a principal in one of those “other” schools described by Mr. Isaac, explained that for many families in her school, poverty was coupled with other,
unmeasured differences that amplify the effect of poverty. For example, like several other principals in the most challenged schools, she explained that the AGT model includes a variable for students who are homeless; however, a high proportion of families in her school did not fit the technical definition of homeless, but were nevertheless in unstable or transitory living situations. In addition, she explained that not only were these children living in poverty, but their basic social relationships were disrupted in ways that affected school outcomes. Many of them, for example, had no parents. They lived in group homes or with relatives other than their parents. Some were children who lacked caring and responsive adults in their lives outside school. These children had less support than children who, for example, had limited means but intact families and stable living situations.

Ms. Dale juxtaposed the strong emphasis she placed on the high quality and engaging instruction with the challenge her teachers had with raising the scores of students, and she wondered if the AGT adequately captured the adversity some students face. Specifically, she described two teachers whom she had given high ratings based on repeated observations but who had low AGT rankings. She wondered if the AGT understated the effectiveness of her teachers in supporting these students toward more positive outcomes.

Similarly, discipline and student crises were palpably greater concerns in the lowest-performing schools. The volume of these incidents in schools with higher levels of poverty seemed to have a cumulative effect on the school independent of each individual incident. For example, principals in three of the lowest-performing schools in the sample were resolving issues related to fights during my interview, and several were
simultaneously juggling crises—even in schools where interviews took place prior to the start of the school year. Also, interviews in low-scoring schools were much more likely to be interrupted—in two cases as many as six times during the interview. This, combined with the principal’s reports about the intensity of needs of their students, suggests that the work of improving learning in some of the most difficult contexts and among the most concentrated poverty is uniquely challenging in ways that are strikingly different across schools. High-scoring schools, in contrast, were peaceful during the interviews.

Principals also clearly struggled with the tension between holding educators and schools accountable for making sure all students learned and achieved at high levels and for punishing educators, particularly in the most challenging contexts, for the deep and systemic adversity that plagued some students. Mr. Grace, the principal of a low-scoring school, explained he felt test scores should certainly factor into evaluations, because if educators do not maintain high expectations for our most vulnerable children, those children will never develop the capabilities they need. However, Mr. Grace also stated that test scores should play a limited role, because there are so many complex dynamics in schools like his that shape outcomes:

I think it’s not fair that you’re going to use test scores when there are all these other issues that affect school. We’re talking about students running off and all that, and threatening to run away. We’ve got to spend time on that, in addition to time on teaching. If a child is saying, “I’m going to run away,” you’re not going to be focusing on instruction at that time. You need to focus on the child’s safety first. And this kind of incident occurs more in some places than others.
Before his children can learn, they need to be safe and in school. In some communities more than others, these are elusive goals.

The unstable and transitory nature of attendance in some schools affects not only those who move, but those who stay, in ways that principals feel are not captured in the AGT. Ms. Island, a principal in one of the other very low-scoring schools, noted that half the students in her school changed schools in a given calendar year. This affected both the continuity of education for students moving in and out, as well as the capacity of the teachers to maintain continuity for students who stay. Teachers were continuously resetting expectations, reteaching routines, and working to reestablish community and familiarity in the classroom community. This is time that is critical to the functioning of a class, but also time not spent on core instruction or maintaining instructional momentum.

Moreover, Ms. Island noted that her school’s low scores incentivize student turnover in ways that compound performance problems. When a school was perceived as a failing school, parents with options tried to avoid it. She asked, rhetorically, “For those of us that are ‘low performing,’ does that mean you’re going to continue to take all of our best and brightest away because our kids are coming in behind already, and we’re playing catch up?” Ms. Island felt that her school’s low scores and the teachers’ low rankings were driven further downwards by the transitory nature of her student population, and in turn, these low rankings exacerbated the problem by incentivizing better-supported and higher-performing children and their families to seek other school options.

Ms. Ivy similarly noted the problem of turnover. Right before the spring tests were administered, she noticed that many new students would show up at her school. Whether or not these new students counted toward the new teacher’s AGT was a function
of how long they were in the new class; most newer students did not count. However, as Ms. Ivy noted, even if the students did not count for the teachers’ AGT rankings, their arrival affected the quality of learning that took place prior to the assessment: “It changes the chemistry of the classroom.” This raises the question of whether, if students were pushed out of one school prior to testing, even if their scores did not count toward AGT rankings in their new schools, the disruption associated with transfers might bias scores downward in the classrooms that were disrupted by new arrivals. Ms. Ivy suspected that some schools might encourage low performers to transfer out, and some schools might systematically be the “receivers” of these transfers. If true, teachers in schools that received these students would potentially have downward bias in their AGT rankings.

Ms. Ivy was careful to challenge the notion that in most cases, one or two students could have an impact on a teacher’s AGT ranking, but she still harbored the concern:

I try to minimize the impact that one child could possibly have on a class when you’ve been instructing them for a year. It’s not going to bring your AGT down to yellow; it couldn’t possibly. Would it? I can’t see that it would, any more than—but do you really have everyone else in your class at proficient is the real question.

Ms. Ivy, like other principals, appeared to be struggling with the tension between her desire to maintain a strong culture of responsibility for student learning on the one hand, and on the other a concern that the tools might not always accurately capture the specific circumstances that led to a given AGT ranking for a teacher. Principals wanted to use the AGT to foster responsibility for outcomes, but using a tool like the AGT for consequential purposes also triggered a sense of responsibility for ensuring that the AGT was telling the “true” story.
Turnover was a concern not just with respect to students, but also with respect to teachers. Ms. Mitchell, a principal in a high-poverty context, noted that her school was still struggling to fill several vacancies in the spring, at the time of the interview. The impact was that many students had spent the entire year with substitutes, because the school could not secure permanent teachers. She observed:

We’ve been pulling together resources using our coaches, using out-of-classroom personnel, using teachers during their conference periods, to still be able to support those students, yet they’re still at a disadvantage. So when they’re looking at scores, you may not have all the data because we don’t have permanent teachers assigned, and those AGT scores are not available.

This lack of a stable, assigned subject teacher was likely to influence the scores of the affected students not only in the subjects with vacancies, but potentially in related subjects as well. For example, if a science position is vacant, students may have fewer opportunities for reinforcement of mathematical skills related to manipulation and analysis of data. And, even if a current-year teacher is not assigned an AGT for those students, those CST scores are likely to be used to generate AGT scores for teachers in successive years. Given that science teachers who incorporate math may support higher math scores, and social studies teachers who emphasize literacy scores may support higher ELA scores, vacancies in those subjects may suppress current-year scores for teachers of other subjects. To the extent vacancies are not equally distributed across schools, this may bias downward the scores in schools with higher numbers of vacancies.

In addition to student mobility, Ms. Dale also cited what she called “huge” variation in class size across the district as another concern. As she explained:
One of the [variables] that is not included is the class size. In our district there are schools that have a class size of 18 to 19 in a class . . . I have classes of 43 . . . so the student here has all the same variables as the student there; the only difference is they’re in a class of 25, and mine are in a class of 45.

Given research on class size effects (see, e.g., Molnar et al., 1999), it is reasonable to question whether a class size of 45 might adversely affect the performance of elementary students in a way that class sizes of 20 to 30 might not.

Most of the examples of “missing variables” described above would potentially bias AGT rankings downwards for teachers for whom these factors were relevant. Most of these variables were described also by principals in low-performing schools, in which principals took pains to stress their commitment to improving outcomes, before they also identified these variables as contributing factors to lower scores and rankings.

Several principals also shared examples of variables that potentially biased teachers’ AGT rankings upward. Ms. Kim led a school in a community with a large proportion of immigrant families whose children were learning English, but whose parents tended to be highly educated. Her teachers’ AGT rankings reflected the ELL status of her students, but Ms. Kim noted that her ELLs were more likely to come from stable, highly educated families than ELLs in some other schools, including some schools in which she had previously worked. She also noted that eligibility for free and reduced-price lunch in many of these families was a function of displacement and geographical disruption, not multi-generational poverty, family disruption, or lack of education. While in some contexts, being an English Language Learner is often a correlate of poverty, lower levels of parental education, and lower performance, in her school, it was often the
opposite. In fact, some parents of her students resented the designation because they felt it held their children back from higher-quality opportunities to learn. She explained that many students in her school were multilingual. For example, some spoke Hebrew, French and English. They might be less verbal in kindergarten, when they first immersed in English, but by first grade, they had caught up or surpassed their peers. As she noted:

They're ready to transition out of being in ELL, even though the district doesn't let them do so until the end of second grade. The parents at that point start to think of the label of ELL like special education. Their children are in the ELL class, which means they can't be in the Schools for Advanced Studies class, a gifted and talented class. This makes the parents upset because their children are not getting access [to opportunities for gifted and talented children] or they're being treated like they're stupid when they're ready to move ahead . . . why are they going to the intervention class when they are the top reader in first grade?

In her context, Ms. Kim noted that her teachers with higher proportions of ELL students were likely to have higher AGT rankings than teachers of ELL students in other contexts, simply because of the highly educated and highly aspirational nature of her parent population, which differed from parents of ELLs in some contexts.

In her context, Ms. Kim wondered if the AGT might actually be sending the wrong signal about the true effectiveness of her teachers by misidentifying certain teachers as highly effective relative to other teachers in the district, when really, a substantial portion of the gains should be attributed to out-of-school tutoring or higher levels of parental support than were available to ELL students in other schools. Although her teachers and school had very high AGT rankings, she suggested:
It is not because we're the greatest teachers. It is because our kids are achieving . . . the reason our kids are achieving is because of what their parents do with them. They came to school with skills. And, I know more than one kid that's getting all kinds of tutoring at home. So, at certain high-performing schools, it's not just a function of the teacher.

Ms. Kim noted that in her previous school, it was not realistic to expect parents to be able to help their children, so whatever happened was a function of the teacher’s efforts, and this made it more challenging to demonstrate gains in student learning.

What the examples above have in common is a concern by principals that some student characteristics or conditions beyond what are captured in the AGT might shape scores in ways that distort the underlying truth about teacher effectiveness. Some were primarily concerned about false negatives: that the measures might identify effective teachers as ineffective. For example, they worried that the AGT might understate the effectiveness of teachers teaching children living in situations of extreme adversity. Others worried about false positives: that the measure might identify as effective or highly effective a teacher who, in their judgment, was actually quite mediocre. In either case, it was the principal’s judgment and knowledge of context that tempered judgments about teacher quality based on the AGT rankings.

What these principals shared was a sense that how the model was structured would shape the way a teacher was ranked, and that the extent to which their own school and teachers and student demographics conformed to the assumptions under the model, their AGT rankings would tell a more accurate or less accurate story about the effectiveness of their teachers. By way of explanation, Mr. Isaac contrasted the AGT

67
rankings with the rankings associated with the growth model put together by the local newspaper, the Los Angeles Times. The LA Times hired a consultant to develop its own model and published its own rankings, using available district data, and the LA Times rankings often differed from rankings assigned by the AGT. Mr. Isaac noted that according to the newspaper growth model, his school was rated “most effective” for two years running. According to the district AGT model, his school was “average, just squeaking by as average on the district’s calculation.” He noted that both models purported to measure the same construct: teacher effectiveness. And both models purported to control for many of the same student characteristics. However, differences in the models generated different rankings. Given this, and given the lack of clarity around why the two models generated different rankings, Mr. Isaac wondered how to reconcile these differences and how to have confidence in these measures as “valid summaries” of the value his school and teachers were adding to the learning of their students. As Mr. Isaac asked, contrasting the results of these two measures: “Then, what is quality or what is effectiveness?”

Inability of the model to separate individual teacher effects from collective teacher effects. The AGT rankings were designed to rank teachers based on their individual effectiveness as teachers. Corcoran (2010) noted, “If value-added measures are to be successfully used in practice to recognize effective teachers, one needs a high level of confidence in the attribution of achievement gains to specific teachers” (p. 5). However, many principals in this sample expressed the belief that perhaps what was being captured by the AGT was not just the performance of the individual to whom the AGT ranking was attached. They described several ways in which the rankings might
actually be better understood as a measure of the collective efficacy of the group of teachers working with the students.

In one school, the principal had very good reasons to assume the AGT rankings were group measures, not individual measures. Mr. Roland was principal of a school that demonstrated substantial gains over the years prior to the study. In his school, groups of math teachers at different grade levels got together at the beginning of the year, figured out who was most effective at teaching different aspects of the math curriculum and students with different levels of skill related to the content, and then shared students so each student was taught a concept by the teacher who was best prepared to teach it to him or her effectively. Thus, although students had a “teacher of record,” the school’s grouping practices ensured that individual students were taught discrete topics by the teachers determined to be the best at teaching them, not necessarily by the teacher of record. Mr. Roland explained that at the beginning of the year he told his teachers:

You're like a parent in that you're responsible to feed them and clothe them and take care of them, but if one of your kids comes in with some complaint about their appendix, are you the best person to operate on them? You're actually doing your job by saying, “No, I’m not, but this person is.”

He structured his teams and schedule to support flexible grouping, and challenged his teachers to share students so that they were distributed to those teachers who would be best able to teach them each particular concept. Mr. Roland went on to add that in the grade levels that piloted this approach, all the teachers had the same AGT rankings, which he described as an accurate reflection of the way the team shared students and pooled strengths.
In Mr. Roland’s school, the response to the AGT was to adopt a logic of greater collaboration, in ways that made the AGT more of a measure of shared teacher capacity. He felt the higher scores over time were a function of shared responsibility for improving learning, and that it did not make sense to talk about individual teachers’ AGT rankings.

In contrast, some other schools opted for strategies that focused on individual teacher accountability. For example, Ms. Island’s school—a very low-performing school with high student mobility—implemented looping in grades K through three, so that children stayed with one teacher, except for transfers between schools, from their time of entry into the school all the way to the first instance of CST testing. This was an effort to heighten the individual accountability of teachers for student outcomes and incidentally, it provided more continuity for students with a lot of instability in their lives outside of school. The cost was that teachers had to prepare to teach a new curriculum every year.

Ms. Island explained that many teachers had complained that their low AGTs were attributable in part to instructional failures by teachers at preceding levels, who had left learning gaps that created a shaky foundation for further growth. According to Ms. Island, the teachers felt and she agreed: “If the child has had inappropriate or lousy teaching before they get to me, and I’m accountable for these scores, that’s not fair.” Ms. Island stated that, because of this belief, and because so many teachers actively tried to move assignments to untested grades to avoid the AGT, she implemented looping in grades K through three to make sure every teacher “owned” the progress of his or her students and every teacher had a chance to be evaluated by the AGT.

Under this new looping approach, the logic went, if a teacher failed to appropriately refer a child for intervention or provide extra help if needed, the effect
ultimately showed up on his or her own AGT ranking, not on the ranking of another teacher. While Mr. Roland’s school used a logic of collaboration and sharing of students to enhance a collective effect on student outcomes, Ms. Island and her school responded to the AGT by restructuring to reinforce the AGT as a measure of individual effectiveness. Teachers in this individual logic had less incentive for professional collaboration.

**Presence of ceiling effects.** Ceiling effects are well documented in the literature on test design (see Ballou, 2009; Briggs & Weeks, 2009; Koedel & Betts, 2010). Essentially, when a test is too easy for a group of students, it has limited value as a measure of what those high-performing students can do and know. When students have perfect scores, they can’t get a higher score if they learn more, and no one simply examining the score can distinguish between the actual level of mastery of two students with perfect scores, or tell which one who knows more than the other.

Two of the principals in the sample led schools that were very high scoring; in each, students often answered every question on the test correctly. These two schools both were aware that being very high scoring posed a unique set of circumstances related to the AGT, and might in some cases make it difficult for very effective teachers to ever rank as anything other than average in effectiveness, or as one principal noted wryly, “gray.” The challenge of very high-scoring children is that if they answered every question correctly the previous year, there is no place for them to go in the current year but down.

Ms. Alana was the principal of a gifted magnet school that was extraordinarily high scoring. She explained: “Growth doesn’t always tell the story . . . when you and
your students don't have room to grow.” In this case, the limitations of the test—the fact that it is not challenging enough to actually measure these students’ current levels of mastery—makes it impossible for it to accurately capture the effectiveness of Ms. Alana’s teachers. Ms. Alana noted that a parent might look at these AGT scores and make the wrong inferences about teacher quality, because the teachers’ instruction was focusing on concepts and skills beyond the basic skills sampled by the test. Ms. Alana said that she and her teachers “put very little value on the AGT here, to be honest,” because they concluded that it just did not speak accurately to the effectiveness of their teachers in the context of the very specialized population they served.

Although Mr. Isaac’s school had a slightly lower mean score overall than Ms. Alana’s school, he described similar effects in his context. The CST measures student performance against a set of defined standards, and many of the students in Isaac’s school had also surpassed that standard. As Mr. Isaac stated: “You’re using a criterion reference test and you’re looking at a population with their bags packed. They might already be up against the ceiling and so there isn’t growth to be shown.” Mr. Isaac reported going to trainings on the AGT and asking:

What if you have the class of students who all score 600s [on the CST]? You have a class of 30 kids that all scored perfectly on last year’s test, but this year, they all missed one question or some of them missed one question. Does that teacher end up being less effective [on the AGT this year, because there is no place to go but down]?

Mr. Isaac also noted that scores are reported in scale scores, and the effect of missing one question on your scale score is much greater for very high scorers than for students
scoring in the middle of the distribution, because the scale is not an interval scale. The test is designed to discriminate among test takers who score closer to the proficiency cut score, so the effect on scale scores of missing one question when you are in the tail of the distribution is much greater than when you miss one question and your score is closer to the middle of the distribution.

This scaling confounds discussions of differences in effectiveness and mastery for the highest-scoring children (and probably the lower-scoring students as well). Mr. Isaac said he had examined the scale scores, and observed that at the top of the scale, missing one question could result in a 50-point drop in scale score, while students scoring near the proficiency cut score would have to incorrectly answer a large number of additional questions before it was reflected as a substantial drop in scale scores. For Mr. Isaac, this raised real questions about what it meant to call a teacher more or less “effective” when comparing teachers who taught students who scored, on average, at very different levels of the test score distribution. As he summarized: “What is ‘effective’ when you have a class of 600s and they all missed one question versus a class where they’re all proficient students and the difference between more effective and less effective is 20 questions?”

For Mr. Isaac, like Ms. Alana, the high-scoring nature of his students, and his understanding of some of the distortions associated with scaling, made him skeptical of the claim that the AGT was a precise measure of teacher effectiveness. The observations of these two principals suggest that in very high-scoring schools, or in classes with very high-scoring students, ceiling effects may distort measures of student growth and in turn, AGT rankings for the teachers of these students.
The interviews in this sample were notable in that, although principals in high-scoring schools identified the presence of a ceiling effect, no principal explicitly discussed the possibility of a floor effect. However, the scaling distortions that occur at the top of the distribution were also plausible at the bottom of the distribution. The CST is a grade-level test. It was likely too hard for some students whose level of mastery was far below grade level. If some of these students were far below grade level, it is possible that even if a very low-scoring student made two years of growth in a single year, this growth would not be captured by the AGT. The closest some principals came to identifying a floor effect is when they expressed a perception that students who score as Far Below Basic are “much harder to move.” It may be that some of these Far Below Basic students have initial levels of learning that are so low and so far below grade level, that even if they make great growth, it would not be captured in a grade-level test. If these lowest-performing students are not randomly distributed across the district, as seems likely given socioeconomic segregation, but in fact disproportionately concentrated in some classrooms and schools, this could disadvantage their teachers in the AGT rankings.

**The AGT as a Relative Measure or an Absolute Measure? What Principals Want**

Because the AGT is a relative measure, we would expect any principal judgments about teacher quality based on the AGT to be explained in terms of a teacher’s performance relative to other teachers. For example, we would expect a principal to describe a teacher with a high AGT ranking as a teacher who is high-performing relative to other LAUSD teachers. However, without some external reference point, it is unclear
whether having a higher or lower AGT ranking makes one a “good” or “weak” teacher in an absolute sense.

In practice, however, some principals spoke as if the AGT were an absolute measure of growth in learning by students associated with a given teacher. This could be because their use of test-based measures was still anchored around the proficiency scores associated with No Child Left Behind. For example, some described situations in which a certain proportion of students scored as “proficient” in one year, and then in the subsequent year, scored as “below basic.” In truth, this pattern could be a function of the placement of the cut score in the second year, and not a function of growth; however, many principals described the teachers of these students as teachers who were not “getting growth,” and expressed their expectation that the AGT rankings would confirm that inference. For example, Ms. Suss described the drop in CST scores and percent proficient she observed as associated with one teacher, and stated:

[The AGT] measures how effective you are at moving children in a certain place and getting to the next point . . . you’re measuring [students’] continued growth and how fast, in a way, that growth is . . . it measures how fast that’s happening. Thus, even when they knew the AGT was a relative ranking, some principals interpreted it in their practice as an absolute measure of growth.

While not all principals said they understood how the AGT model worked, they all appeared to recognize that a teacher who was able to “move” her students or improve learning a lot, could potentially receive a higher AGT ranking, even if her students were all scoring as “basic.” As Ms. Rosa explained:
You’re looking at the students’ growth over the time that they’ve been with that teacher. So if I’m an English learner, I was an English learner before I came in. If I’m a gifted student, I was gifted before I came in. And this is where I started and this is how much growth I made over the time that I was with you. So it was a way to measure the impact of that teacher’s instruction on that child in a year’s time and to do that over multiple years not just one.

This emphasis on growth was said by principals to be more “fair.” For example, Mr. Isaac understood the AGT was modeling data in a way that actually might shed light on how educators in any school could evaluate the relative effectiveness of their efforts to improve learning. In the lowest-performing schools in the sample, the principals explained that, because their students came in scoring lower and stayed lower-scoring overall, these schools were disadvantaged by test-based measures that focused on student scores, rather than the relative success of efforts to improve those scores. For them, the AGT represented a chance to honor progress made by their students and teachers, even if they were still relatively low scoring.

Overall, while principals recognized and appreciated the emphasis of the AGT on growth and learning, many also identified factors exogenous to the AGT model that might bias the AGT rankings either systematically (ceiling and floor effects, missing variables) or for individual teachers (nonrandom sorting) in ways that might lead the AGT to misidentify individual teachers as more or less effective than their colleagues. Some principals were concerned in turn that this misidentification could lead to perverse incentives that would erode their efforts to improve student learning and opportunities. For example, teachers who were falsely identified as lower ranking might be discouraged...
from teaching the students who would most benefit from their skill or otherwise demoralized. In contrast, teachers who were falsely identified as high ranking might be positively reinforced for practices principals wanted to discourage, like didactic teaching or transferring of challenging students from their classes. Either of these scenarios potentially undermines efforts to improve instruction.
Chapter 5: The AGT as a Lever to Improve Practice or a Hammer that Erodes Practice?

Principals’ understanding of what the AGT rankings measure, as well as their confidence in their own understanding, shaped how they used (or did not use) the AGT rankings in their own professional judgments about a teacher’s efficacy, as well as how they used (or did not use) the AGT rankings in their efforts to improve instruction.

Many principals did look hard at AGT rankings, and in several cases, said that AGT rankings caused them to reassess their previous assumptions and evaluation ratings of certain teachers, and which thus led them to reassess the focus and direction of their supervision. Other principals, where the data were available, used AGT rankings for subgroups of students (e.g., poverty status, race, disability, English language status) to inform conversations about which types of students were more or less supported by a specific teacher’s instruction. And most compellingly, principals in some of the low-scoring schools used the AGT rankings assertively and confidently to leverage change and create a sense of urgency among teachers about educational improvement. Because the AGT rankings were summative in nature, and because in many cases, subgroup data were not available for the teachers observed by principals in the sample, the AGT was described as a somewhat blunt tool whose primary benefit was to motivate improvement. However, across the board, principals did not describe it as a tool that helped them understand how teachers needed to improve, or what specific feedback to provide to ensure better outcomes for children.
Several patterns in the sample of principals merit attention and will be discussed in this chapter. In particular, there were notable differences in how principals in the highest-scoring schools described the AGT rankings, and how principals in lower-scoring schools described the AGT rankings. For example, none of the principals in the higher-scoring schools was inclined to give the AGT ranking much weight in teacher evaluations. In fact, three of them discounted the AGT ranking, implying it might incentivize poor teaching practices and suggesting that parents of their students were also not highly concerned with the AGT rankings. These principals gave more weight to other measures of effectiveness when assessing teacher efficacy, such as their own judgment or other sources of evidence, including classroom-based work samples. In contrast, principals in low-scoring schools were more likely to report using AGT rankings in conversation with teachers to create a sense of urgency or responsibility for improving the learning of students.

Most principals expressed the belief that while they found looking at the AGT data was helpful, they should not be used in the absence of other data, or replace professional judgment. As Mr. Thom stated:

I think it's always good to have more information in evaluating the teacher, but [the AGT] shouldn’t be the only measure. As to how much it should count, I think it should be used as one of many tools. And if you are thinking percentage-wise, I would probably not put too heavy of a percentage of it.

With few exceptions, these principals suggested that test-based measures needed to be used in context and in conjunction with other measures.
In this chapter, I discuss specific ways principals described using, or not using, AGT rankings in supervision and evaluation.

**Using the AGT to Identify Strengths and Weaknesses of Teachers and Target Improvement Efforts**

Schools have limited time and resources. To improve instruction, they need to be able to target their attention and resources at students and subjects and grade levels most in need of improvement, in order to achieve the greatest benefit for students and the public overall. One of the greatest benefits of the AGT, as identified by principals, was the potential it offered for more effective targeting of improvement efforts.

For example, Ms. Ivy noted that in her school, the leadership team had historically decided which teams and grade levels were weak or strong based on the percent of students that scored as proficient. Over time, however, as they examined the data, members of her leadership team realized that the tests in some subjects and grades seemed to be comparatively harder than the tests in others. Changes in “percent proficient” year-to-year and subject-to-subject seemed to be a function of where the threshold was set, as much as the progress of the students. If a cut score was relatively close to the performance level of the majority of students, then the school observed a bigger increase in percent proficient on that test. If, at the next grade level, the cut score was relatively higher compared to the average performance of the students, a lower proportion of students would score as proficient, even if other school-level assessments and benchmark assessments suggested students were making a lot of progress. This in turn might lead administrators to misdirect attention and resources towards teams or teachers that were
making progress but were assessed by harder tests, and away from teams or teachers who were actually stagnating, but for whose students the tests were relatively easier.

However, the AGT rankings challenged Ms. Ivy and her team to re-examine assumptions and shifted Ms. Ivy’s understanding and the school’s internal conversation about improvement priorities. The longstanding assumption in her school had been that instruction in English must be stronger than instruction in math, because the English department was closer to achieving its performance targets. However, the AGT rankings suggested that the math department was getting greater academic gains relative to other teachers in the district and compared to the English teachers. As Ms. Ivy noted:

[The math department] was having tremendous growth, and just need[ed] to capitalize on what they were doing so they could continue the growth. Even though their scores were higher, the English department was stagnant, so that’s where we really had to go in and figure out what we needed to do to restructure the work so we can actually achieve growth.

To the extent the AGT rankings accurately signal which teachers and practices could generate greater relative gains, it could guide the efforts of leadership to target scarce resources at teachers and subjects where gains were weaker relative to others in the district.

Similarly, several principals noted that every year a larger proportion of students in their schools scored as basic on the math CST in fifth grade relative to math in other grades. As a result of this pattern, many had assumed their fifth-grade teams were weak in math. However, when the AGT scores came out, some of these principals reported revisiting this assumption, because all of a sudden, they were able to compare the growth
their teachers “added” relative to the growth “added” by other teachers at the same grade level. As Ms. Rosa explained:

Those dismal fifth-grade math scores were average in the district. That made me think about the teachers from the school that I came from previously [whose fifth-grade students had also appeared to “slip” to Basic].

Several principals who noticed this pattern hypothesized that the fifth-grade math test might just be a harder math test for students overall, and that what principals needed to focus on was which teachers were getting improvements in learning in each subject, relative to their peers, because these teachers seemed to be more effective than others regardless of the difficulty of the test. Students in “easier” subjects or grades might have more students scoring as proficient, but might not actually be making progress. Others questioned whether they needed to revaluate math instruction district wide. This moment of reflection, which was triggered by juxtaposing proficiency rates with AGT scores, shifted principals’ focus from individual teacher weaknesses to broader causes of low performance relative to goals.

The AGT as a Lever to Improve Practice

Several principals reasoned that if district and school leaders can use the AGT to adequately measure the specific impact of individual teachers on outcomes, they can effectively target efforts to improve instruction at teachers who need improvement. And, they can identify, support, and sustain teachers who effectively achieved results. If class sizes and subgroup sizes are large enough to provide reliable measures and different AGT scores can be generated for different subgroups within a class, principals can use the
AGT to identify whether teachers are more or less effective with different subgroups, and target professional development accordingly.

Principals in the sample observed that, to the extent that the AGT models for demographic differences that shape performance, it removes one of the excuses some teachers might use to resist making improvements in instruction. Many principals, particularly those in low-performing schools, suggested the AGT rankings were powerful as a tool for challenging teachers to take responsibility for the learning of their most disadvantaged students. Because well-done teacher observations take time, it is difficult to thoughtfully observe every teacher on a large faculty in a timely way. However, the AGT rankings were generated remotely by the district and provided data for every teacher. Thus, AGT rankings potentially served as a relatively inexpensive and accessible, theoretically objective summative assessment of each teacher’s effectiveness relative to other teachers.

AGT rankings appeared to promise principals a tool to easily audit effectiveness and hold teachers accountable for improving outcomes. However, for principals in this sample, the value of the AGT as a tool to improve teacher performance was tempered by two main concerns. First, as noted above, although the AGT appeared to rank teachers by their effectiveness, it actually provided principals with very little—if any—information about how individual teachers could improve. And, as implied in the preceding chapter, principals had some concerns about whether the AGT adequately measured the learning that matters. Specifically, they were worried test-based measures might reward test-focused instruction that raised scores, but that did not provide students with the breadth of opportunities they would need to succeed in more advanced learning.
When principals described the role of the AGT in improving practice, it tended to be in the context of supporting evaluative judgments, including in the context of dismissing teachers. As Mr. Roland explained:

I’m not too sure that AGT would have the same importance put on it if it was easier to say, “You're not working to the level you should be.” There are states where, if a teacher is just not doing a good job, at the end of the year, you cut them loose. If you follow due process, then they're gone. In this state, it is much longer than that. You look at LAUSD housing hundreds of people who should not be in education, should never have been in education.

In this context, test-based outcome measures potentially serve as additional, “objective” information that can be used to provide feedback to resistant or underperforming teachers, and if they fail to improve, move them out of the profession.

For the most part, principals tended to present test-based measures as external evidence to inform their assessment about effectiveness. Principals across all contexts agreed that the AGT put pressure back on teachers to evaluate the contributions of their own actions to student learning, but noted that technical limitations weakened the value of the AGT as a tool. Beyond that, principals either emphasized the AGT as a tool for creating urgency and targeting instruction, or they stressed improving teaching and learning, and consulted the AGT as one more piece of data to cast light on the relative effectiveness of improvement efforts. The former, more “urgent” approach was more common in the lower-performing schools, while higher-performing schools tended to focus on a broader definition of improving instructional practice than that captured by test scores.
Use of the AGT by Principals in Lower-Scoring, Higher-Poverty Schools and in Higher-Scoring, Lower-Poverty Schools

Principals in lower-scoring and typically higher-poverty schools were more likely to describe the AGT rankings as a tool for motivating teachers to improve teaching practice. These principals tended to frame accountability in terms of ensuring better test scores. The logic in these contexts was that if the data suggested a teacher was relatively less effective at supporting academic growth than colleagues, this would be an opportunity to sit down, perhaps with the Teaching and Learning Framework, to discuss how a teacher could be more effective.

In addition, some principals described situations in which the emphasis on using AGT data to improve performance might lead to a narrow realignment of curriculum and instruction around test-based performance goals. For example, principals described using an analysis of benchmark and summative test data to focus instruction on material that mattered on the test or on which students did not score well on the test. Mr. Frank said he used the AGT to identify relative weaknesses and prioritize efforts to improve practice. He described a teacher who had focused intensely on reading comprehension, but had not paid as much attention to math. “I let math kind of slip away,” the teacher told Mr. Frank after learning about his students’ test results. Mr. Frank used this opportunity and the AGT rankings for each subject to motivate the teacher to think more analytically about the way she taught math to get a better understanding of areas of strength or gaps and areas that needed work.
For Mr. Frank, the challenge was to create a professional culture where adults felt safe looking at outcome data and getting feedback, and using insights from this review to improve their work. Mr. Frank described the need for a culture in which teachers felt safe acknowledging the need for help or support—one in which “it's okay to open your classrooms to other people to support and help you, to raise your AGT score, but at the same time just increase student achievement.”

Ms. Ivy also described the AGT results as providing an opportunity for reflection about what next steps teachers needed to take to improve their practice and in turn, improve outcomes. Her elementary teachers could look at results for two subjects and frame their professional goals in response to priorities suggested by the data. When teachers had a lower AGT ranking for one subject than another with the same students, Ms. Ivy said, teachers needed to ask:

What is it I need to do differently for math? Is it that I need professional development? Is it that I need to collaborate more with others? Is it that I’m not using appropriate strategies? [The AGT] really begins that reflection period. It is important to know that when the children left your classroom they were better off as a result of your teaching than when they walked in.

In theory, this could lead to more effective targeting of scarce professional time and resources, especially if areas of lower “growth over time” were also areas of lower overall performance.

In Ms. Mitchell’s school, which was lower-performing, leadership reviewed AGT rankings with individual teachers and used rankings to set priorities for the year. She also noted that she used the AGT to flag individual teachers for monitoring. What was
interesting about Ms. Mitchell’s use is her implicit operating assumption that the AGT is an absolute measure of growth, and that growth can be compared across subject areas. In fact, she seemed to use the AGT and the students’ CST scores almost interchangeably:

We have about six teachers whose AGT scores over the past three years have not grown, and in some cases have declined. And one in particular that I work with directly, the same students are cored [paired with] with another teacher for a different set of subjects, but their AGT is moving forward [in those other subjects]. . . . Why are the scores not going up for this group in one subject, [when they are going up for the same group with a different teacher in a different subject]?

In Ms. Mitchell’s view, the AGT scores became a powerful lever for improvement, because she used them explicitly to hold teachers accountable for student performance.

In contrast to the higher-performing schools, where principals tended to take a nuanced view of what the scores meant and what might shape the scores themselves, in Ms. Mitchell’s lower-scoring context, the teacher was held personally accountable for moving that metric. Ms. Mitchell described education in industrial terms:

We’re producing a product. And if the product is education, there needs to be some accountability for the product that we’re producing. Too many times we have students that are not receiving quality education and yet the person that is responsible for that is still going to be employed and paid. The argument [against accountability] is that not everyone gets the same type of students. But everyone is a teacher, and so you need to be able to utilize the resources available.

By holding teachers and themselves accountable professionally for improved test
performance and discounting responsibility for the students, principals like Ms. Mitchell create a sense of urgency around raising test scores. Because the AGT model was structured to account for factors like poverty, language learning, and disability status, Ms. Mitchell said the AGT removed those as excuses for poor results. By using the AGT, she was making it clear that she expected her teachers to do everything they possibly could to improve outcomes.

Similarly, Ms. Irma felt the scores, though perhaps not an accurate measure of effectiveness in every case, did capture useful information overall about the effectiveness of her teachers. She too wondered at first if the AGT scores were fair, but after reading more about the rankings she concluded that the science behind them was sound. Like Ms. Mitchell, Ms. Irma was confident enough in the AGT scores to use them to push teachers to be more accountable. For example, she sat down for a conversation with an English teacher, whose ranking was on the “lowest end possible” of the AGT rankings. He offered a number of reasons by way of explanation: that he taught the Developing Readers and Writers Course (a secondary literacy intervention curriculum for low-performing students) and that he had the most students. Ms. Irma said that to get him to take ownership for improving students’ learning, she shared the description of the AGT model that identified the student factors for which the model accounts, and said:

Come on! We have to be honest here. You can’t blame my kids any longer. You can’t blame that they’re Hispanic. You can’t blame that they’re poor. You can’t blame that they’re homeless. You can’t blame all these factors. Instead maybe you need to turn the mirror around.
In sum, principals in low-performing schools like those of Ms. Irma and Ms. Mitchell sometimes used the AGT as a tool to push teachers to take individual ownership for improving results and to direct and target their improvement efforts.

Some principals didn’t trust the AGT on its own as the basis for consequential decisions, but were motivated by the AGT to pay closer attention to the specific outcomes of the students of certain teachers. When a teacher had a relatively low AGT ranking, these principals went back to other data points for that teacher and her students, to develop a more robust picture. For example, Ms. Brown explained, she didn’t just take for granted that, because a teacher received a low ranking, she was ineffective. However, for her, when both the AGT ranking and her subsequent observations of the teacher’s instruction appeared to support the same judgment about teacher quality, the AGT served as powerful confirmatory evidence.

Ms. Rosa stated that she did not use the AGT much, but she noted that having data from a test-based measure did add some weight or urgency to the conversation about improving learning. In effect, the AGT provided another data point to cross reference against other evidence on teacher effectiveness. As Ms. Rosa said: “It gives you more, I don’t want to say ammunition, but more substance, more substance for you to have those data conversations with the teachers.” Having the numbers focused her conversations with teachers away from whether or not they were getting results, to questions about why their children were or were not demonstrating more progress.

Not all principals in lower-performing schools were so confident in the AGT measure, and some noted that at times they were not sure it aligned with what they expected. Ms. Dale, for example, described a handful of teachers who, based on her
observations, were highly effective, but whose teaching “just didn’t translate” into relatively higher AGT scores. As an administrator, she was concerned these low rankings could demoralize teachers who, in her judgment, were working hard and doing a lot of the right things in their classrooms. In addition, she felt that many of her teachers were already trying new strategies by the time the AGT rankings arrived. Ms. Dale explained: “Bottom line, it’s a measure. It tells you something. But bottom line is you have to make your plans and move forward, so I’m not always hitting them over the head with their AGT.” When the ranking was low, she responded by moving forward and focusing on current data and current students, rather than emphasizing low rankings based on prior-year students.

Not all principals viewed AGT as a lever for improvement. Noticeably, the principals in the higher-scoring schools were more skeptical of the AGT as a metric for effectiveness, and placed less emphasis on it in evaluative conversations. They were not defensive about the AGT; instead, they were reflective about its technical limits in their specific contexts, and thoughtful about how, in those contexts, they could best level improvement. These principals reported that their teachers felt similarly. As Ms. Alana said: “Now that [the AGT scores] are out, they use them. They reflect on them, but in some ways they do not feel like they tell the whole story.” Ms. Alana went on to note that because so many students in her school had perfect scores, there was really no way for them to demonstrate improvement, so it would be unlikely that her teachers’ AGT rankings would ever identify them as anything better than average, because of the ceiling effects described in the preceding chapter. In addition, the teachers in her school who had lower-scoring students and lower AGT rankings were also teachers who Ms. Alana
perceived as serving disproportionately challenging students. For example, she described a scenario in which a specific teacher had a lower AGT score, but also served a disproportionate number of kids struggling with emotional issues, because she was widely recognized as effective with these children. In Ms. Alana’s judgment, students with emotional issues, while identical demographically to many other children, were less likely to show progress. Thus, Ms. Alana saw the teacher’s lower AGT ranking not as an indicator of lower teacher efficacy, but rather, an indicator that teachers who disproportionately serve students with serious emotional challenges are more likely to have lower AGT rankings. Given this reality, Ms. Alana felt strongly that the AGT could not be used for evaluative purposes, and thus should not be relied on as a driver for improvement—unless it were used in conjunction with professional judgment.

Higher-scoring schools were also more likely to have a diverse student population, because they were less likely to be uniformly high-poverty, Latino, or African-American, and less likely to have very large class sizes. One paradoxical effect may have been that this diversity made it less likely that these schools received reliable estimates of teacher effectiveness with different critical student subgroups. In theory, the AGT could provide principals and teachers with data about teachers’ relative effectiveness with different subgroups. In practice, however, this subgroup analysis was not always possible, because the numbers were small or measures were unreliable for subgroups, so data were suppressed.

Mr. Isaac, a principal in a high-scoring school, was interested in seeing if the AGT could provide insight into whether his teachers were more effective with some kinds of students than others. However, Mr. Isaac noted that, because subgroups in his
school were not large enough to support reliable AGT scores for different subgroups, this kind of analysis was not possible: “In all cases it came back as not enough students to validly report.” Thus, the use of the AGT for improvement purposes was somewhat circumscribed by the technical limits of the tool.

In higher-scoring schools, principals may also have had more freedom to discount the AGT, because parents were less concerned about performance as captured on the CST. Ms. Alana noted that parents in her school were more focused on other indicators of whether their child was happy and intellectually engaged, and did not seem focused on the AGT. As she said: “I don't think I’ve had a single parent come into my office and say, ‘You know I’m concerned because my child’s teacher’s a red [Far Below Predicted].’” In this high-performing school, Ms. Alana viewed test-based measures as just one data point in a holistic assessment of her teachers’ effectiveness. And, apparently, her parents did as well.

Ms. Kim, a principal in a higher-scoring school, did not give the AGT much weight in her evaluation of teacher effectiveness, nor did she use it explicitly to create urgency in teachers to improve practice. Like Ms. Alana, Ms. Kim treated the AGT as just another data point. Ms. Kim also articulated that she was disinclined to emphasize the test because to do so was inconsistent with her values with respect to effective teaching. As she explained:

It's interesting because at this school, they're not super focused on teaching to the test, which I like, which goes with my philosophy because it's about teaching and then at the last minute you go in and hit those skills again. . . . You test-take for practice and you do review, but mostly it's just about good teaching.
She contrasted that with a previous higher-poverty school where she had worked. That school was very focused on test scores and on teaching in specific ways that enhanced test performance. As Ms. Kim explained: “I would hate for them to have a teacher that taught to the test because life would be really, really boring; and I want them to be taught to the whole child.” Ms. Kim expressed the concern that, while some instructional strategies might boost scores, they might also disengage students or narrow school experiences in ways that had other negative consequences. Comments like those of Ms. Alana and Ms. Kim raise questions about whether teachers in higher-scoring schools feel less pressure to teach to tests and demonstrate test score improvements, in ways that may lead to substantive difference in the breadth and depth of learning opportunities to which students are exposed in higher- and lower-scoring schools.

Mr. Isaac, a principal in a high-scoring school, discounted the AGT for entirely different reasons. He viewed the AGT as a relative ranking and not an absolute measure of the growth associated with a certain teacher’s efforts. This made it difficult to know how to use it in any meaningful way, because there was not an external standard against which to assess the meaning of the ranking. As he saw it, the AGT simply told a teacher that he or she was “more than effective district-wide,” but did not necessarily say whether the teacher was an effective or ineffective teacher.

Mr. Isaac understood that the AGT was a relative ranking at best. Assuming the model accurately accounted for all factors that shape student growth and learning (which Mr. Isaac did not assume) the model told him how much score improvement was associated with a given teacher, compared to other teachers of similar students. Mr. Isaac pointed out that in the absence of any external benchmark or standard, this kind of
ranking was relatively meaningless: in any kind of ranking, there is a distribution. Someone has to be best, and someone has to be worst. What the AGT did not tell him was how skilled the group was overall: were the teachers being compared all Olympians of teaching, such that even the lowest performer was still extraordinary? Or were teachers district-wide so weak that even being better than average did not make a relatively effective LAUSD teacher competent in an absolute sense? Without some external benchmark for effectiveness, Mr. Isaac was skeptical of the AGT rankings, and all the more so when he considered some of the factors that might distort the rankings, such as ceiling effects on children’s scores in his high-scoring school. Relative assessments of “effective,” in Mr. Isaac’s mind, are only meaningful if we know how “effective” the group as a whole is relative to some external standard, and whether performing better relative to the group as a whole actually leads to better outcomes for students over the long run.

Mr. Isaac also noted that the color coding and normative language associated with different levels of ranking appeared agenda driven, especially as it was not yet correlated in any predictive way with any substantively important external variable such as success in future advanced coursework. A teacher could be labeled green or blue for being relatively effective, even if she were weak in an absolute sense. In contrast, a teacher could be labeled yellow (below average) due to anomalies of testing and ceiling effects that made it hard for her students to demonstrate a lot of growth.

Mr. Isaac did not dismiss the potential value of having some test-based measure of teacher effectiveness; in fact, he sought one. He was just skeptical that this particular measure was sufficiently robust to do what it purported to do. As a principal, he
considered and reflected on test-based measures, but did not use them to drive decisions in a singular way. For example, he described looking at growth rankings of teachers who applied for positions in his building, and thinking about them in the context of the populations and schools in which these teachers taught. However, he made his decisions to hire on the basis of other factors.

Despite his skepticism, Mr. Isaac did note that the AGT rankings, if used carefully and appropriately, could have the effect of creating peer-to-peer pressure to improve. He noted some teachers had individualistic mindsets of teaching, and thought: “If I’m doing a good job in my room, but the person next to me is doing a crappy job, as long as the crappy job they’re doing isn’t impacting me, their kids aren’t running through my line at lunchtime and disrupting my whole thing, it’s well, it’s none of my business if they’re doing a crappy job.” He contrasted this individualistic mindset with a more collaborative, team-based mindset of teaching in which individuals hold each other accountable for educating students well. In this more collaborative culture, teaching would be a team effort where people start saying, “Hey, well, if you don’t pick up the slack and you’re not going to carry your weight, we have a problem with that.” Viewed through this lens, the AGT rankings could be used as a tool to build the collective investment of teachers in group outcomes, especially if used as a school-level measure. This observation potentially has powerful implications for how principals lead and manage their schools’ professional culture.

Mr. Roland’s school was not high scoring, but it was a school that had made substantial gains in recent years, particularly after Mr. Roland led his faculty through a process of identifying which standards were given the greatest emphasis and were
associated with the most test points on the CST. Mr. Roland explained that the AGT could serve the purpose of targeting interventions and establishing urgency. However, he stressed that improvement comes from the professional supports to which the teacher has access, and not from the AGT itself. As Mr. Roland stated:

I think [the AGT] opens the conversation. It’s better than what we had before. It’s shifting from saying “I have a kid whose mom is in prison, whose dad is in Mexico. I can’t do a thing with this kid,” to maybe, “Wait a minute. This kid’s mom is in prison, his dad’s in Mexico—what can I do to help support this kid?” It is shifting to “Maybe I need to start building up his self-esteem,” rather than saying, “Yeah, too bad, you'll never amount to anything.”

In sum, Mr. Roland saw the AGT as creating an opportunity to have a conversation with teachers about the fact that no matter what challenges children bring to school, it is the teacher’s job to move the child’s learning as far as possible.

Mr. Roland recognized the tension between holding teachers responsible for overcoming life circumstances on the one hand, and letting teachers off the hook for trying to improve student outcomes, despite life circumstances, on the other hand. The AGT rankings sit right at the intersection of those two impulses. Despite this, like many principals, Mr. Roland indicated that if he worked hard on developing his teachers and they worked hard on understanding and challenging every child, the scores would take care of themselves. His focus was on developing his teachers’ practice.
Aligning Instruction to the Test

Several principals described teachers’ responses to using test scores for the purposes of teacher evaluation, some of which might bias the AGT as a measure of teacher effectiveness. Specifically, like Mr. Roland above, principals were aware that very closely aligning what is taught to what is tested, as well as to the format in which content is tested, is likely to raise scores.

Some principals recognized that the test prioritized some content in the standards over other content in the standards, and thus, teachers who emphasized the tested standards, as opposed to standards that were in the framework but not emphasized on the test, could raise scores, even when they were not addressing the full framework of standards. As Mr. Isaac explained:

There are release questions that come out from the state that our teachers look at and our parents are aware of the tests. One of our priorities is making sure the kids are ready for the test and we have the periodic assessments, which are correlated with scores on the CST. Teachers know what the standards are, and what the blueprint says about which standards are tested more than others. Teachers spend a lot of time looking at the academic language that are in those release questions to make sure that students are familiar with vocabulary that they’re going to count on the test.

If teachers teach to the heavily sampled standards and to the specific language and formats evident in the release items, they may give students a score boost. Mr. Isaac added that teachers could be incentivized to teach narrowly for the test in this way, if their names and associated AGT scores were going to be published in the L.A. Times. He
noted that by the time teachers have administered the test several times, they have become familiar with the test to the point where they have memorized some of the questions. Teachers also began to identify which aspects of the test caused the greatest challenge for their students, and targeted instruction at particular content that students were likely to encounter on the test.

Some principals described looking at score reports to identify priorities for changing instruction. Some of these priorities had to do not with whether students understood concepts, but with whether they were familiar with specific ways these concepts would be sampled on the test. For example, Ms. Suss used the example of syllabication to explain how it was not enough to teach a concept; teachers also needed to be attentive to the format in which the concept was sampled on the test, and then fine tune instruction to address tested formats. In her school, for example, teachers tended to teach syllabication by breaking words into parts and having students clap to emphasize the syllables. However, on the CST test, Ms. Suss explained, words are broken into word parts with a hyphen in between. She realized her students could identify syllables when asked to clap them out, but the test didn’t ask them to clap. Her students were not making the connection between the auditory strategy they used in class and the visual strategy called for on the test. Ms. Suss’s example suggests that what matters is not just whether students know a concept, but whether they know the specific way the concept will be assessed on the test. This suggests that a second way to improve scores is by teaching to specific formats used on the test for accountability purposes. Students who practice with specific formats will perform better, whether or not their conceptual understanding is better.
Mr. Roland’s teachers also closely analyzed the test, not for test formats, but to identify which standards accounted for the greatest proportion of points on the test. They reasoned that the most heavily tested standards were the most important standards, at least for the purposes of accountability. In part, this reflected their belief that the standards were expansive, and teachers did not have time to teach to mastery on all standards covered by the test, especially given their student demographics. Instead, they effectively did educational triage. Mr. Roland and his teachers audited the test to identify those standards that were associated with the greatest number of test questions in any given year, and determined that those were the most important standards to teach. For example, he said, “If you’re talking about math, and you’re talking about common fractions . . . and they have five out of 20 questions on adding mixed numbers, then what they’re saying is, ‘This is the most important standard.’” In other words, the way the standard is sampled on the test, and how heavily sampled it is, ends up defining what it takes to “master” a standard.

Those standards and the specific ways they are sampled then became the focus of instruction. As evidence of the effectiveness of this strategy, Mr. Roland noted that in the year they made this shift, the school had a 50-point jump in its API score. Mr. Roland explained: “The rationale I've given teachers here who say, ‘Well, what, you want us to do is teach to the test’ is—‘No, we're teaching to the standards that have the most emphasis from the state, because they give more questions. They really want to know these kids have these standards.’”

While narrowing instruction to tested content and formats will improve tested outcomes in the current year, it may adversely affect longer-term outcomes. If some of
the grade-level content was not emphasized on a given grade test but was foundational for learning in subsequent grades, gaps in instruction might affect the scores of subsequent teachers, and not those of the teacher who omitted that instruction. As Ms. Charles explained:

If a third grade teacher does inflate scores or results and then kids go to the next grade level and the next teacher no matter what they do, if they’re teaching with fidelity but the kids now fall down [due to gaps], yes, it’s going to reflect on their AGT. It’s going to look as though they didn’t bring these children up to the level that they should.

This raises questions about whether AGT scores for some teachers at higher grade levels could be biased downward if their students were disproportionately taught in previous years by teachers who taught narrowly to content emphasized on the test. Given that principals in higher-poverty, lower-performing schools felt the greatest pressure to improve scores, any such narrowing could be particularly problematic for their students and teachers.

Like Ms. Charles, Ms. Ivy worried that when teachers are ranked based on their students’ scores, they might sometimes teach in ways that are good for rankings but perhaps not good for students. As a middle school principal, this was of particular concern, because the preparation in the elementary schools that her current students attended before middle school influences how well those students thrive in their current middle school setting. She noted:

My fear is that teachers will begin to do things that may not necessarily be in the best interest of instruction as a whole in order to make sure that their kids are
scoring better, in order to make sure that their AGT shows the growth over time that the kids are supposed to have. I was talking to a colleague of mine who was telling me, “We keep on getting kids from this school, and the kids are coming in scoring as advanced, but when we test them mid-year, they’re not advanced.”

Ms. Ivy was concerned that some teachers might teach in ways that enhanced their AGT rankings, but did not leave their students well prepared for more advanced work in subsequent grades. By way of example, she stated she had recently hired a teacher in part based on the teacher’s high AGT ranking and demonstrated effectiveness in achieving gains with English Language Learners. However, when she actually observed in the teacher’s classroom, what she saw did not line up with the high AGT ranking. Ms. Ivy said this raised questions for her about what the teacher had done in his previous school that had led to the high scores and high AGT ranking, and specifically, whether something unethical was done with the tests or whether he spent the entire year prepping for the test. She explained that when she observed his actual teaching, she was compelled to question the AGT ranking on which she had relied.

Mr. Isaac pondered the increasing stakes associated with incorporating test-based measures into teacher evaluations, similarly wondering what impact it might have on instruction. For example, he noted that the learning described in the standards seemed much broader than the content emphasized by the test. In particular, he felt the tests emphasized content knowledge, while the standards emphasize inquiry and scientific thinking. Mr. Isaac noted:

There are a lot of knowledge level expectations on the fifth grade science CST. Meanwhile, best practices and science education call for inquiry practices, getting
kids working hands-on and working collaboratively and constructivist inquiry-based kinds of things also matched up with the content. I know our students excel at that because they go to science lab every week, and I’ll go into science lab and I watch them and they work together and they solve problems. They have amazing discussions and ask amazing questions about the science concepts that they’re working with, but all of that experience which is so important in my mind for science education is not really measured at all on the test.

For the most part, Mr. Isaac says his science department follows the district curriculum and state curriculum, but commits to these best practices, and for the most part, he feels his teachers have been able to teach well and still achieve strong improvements in learning. However, his comments reflect his concern that test-based evaluation measures might reshape his teachers understanding of what is important learning, and limit it to what is easily tested.

Several principals identified this tension between doing what works for raising scores and what works with respect to engaging children broadly and deeply as learners. Ms. Cook explained:

As a parent, too, it's like I know what I want my kids do on the CST because then they'll qualify for this, that, and the other thing; but I would hate for them to have a teacher that taught to the test, because life would be really, really boring; and I want [teachers] to teach to the whole child.

Like Ms. Cook, several principals identified types of learning opportunities they thought might get squeezed out by test-focused teaching, including an emphasis on inquiry and opportunities to connect classroom learning with real-world experiences—all learning
opportunities that might be more engaging, but which are less directly connected to test performance. At the same time, however, these principals expressed the belief that the test scores were capturing some critical learning, and that improving performance on tests was a priority.

This tension was a conundrum that surfaced in many interviews. Principals were committed to improving learning, and confident that if they and their teachers did, scores would rise. At the same time, numerous principals expressed concern that some teachers with very high AGT rankings did not, in their observations, appear to be teaching effectively in engaging and challenging ways, and that some types of instruction might yield higher scores in the short run, but at the expense of breadth and depth of learning and student engagement with the subject. In other words, improving learning was a priority, but the AGT rankings would not necessarily distinguish between teachers who improved learning and teachers who improved test scores, and the difference mattered to most principals.

If teachers can raise their AGT rankings by teaching narrowly to the test, as these principals suspect, the AGT may inadvertently direct teachers toward practices that raise scores by narrowing the breadth and depth of student opportunities to learn in ways that may disadvantage them in other contexts beyond testing, and potentially even in subsequent years. This would erode the value of the AGT as a measure of teacher effectiveness. Ms. Richmond noted this unintended consequence when she observed that if scores were inflated for one year, the AGT ranking for a teacher in a subsequent year would probably drop.

In addition to concerns about the effect of test-based accountability on the
substance and quality of instruction, some principals expressed the concern that pressures of testing were driving movement within schools that undermined efforts to support stable, continuous improvement of learning. For example, the comments of several principals raised questions about whether the AGT rankings were driving increased mobility by teachers across grade levels within schools. Several principals described efforts by teachers to seek reassignment to non-tested grades, or movement across grade levels within schools. Of note, one of the statistical properties of the AGT is that the model seeks to increase precision by using data from prior testing years. Teacher movement across grades increases the challenge of developing reliable AGT measures for individual teachers. It also undermines efforts to develop stable strategies for improving learning at each grade level. This challenge was aggravated in some places as enrollments declined, forcing disruption and reshuffling of positions across grade levels.
Chapter 6: What Do Principals Think is Measured by Standards-Based Observations Based on the Teaching and Learning Framework?

With respect to the AGT, I argued that for a tool to be a useful measure of teacher effectiveness, it needs to measure what stakeholders agree is indeed effectiveness, and it needs to do so reliably and accurately. The same standard applies to the TLF. While the AGT relies on outcomes (measured in test scores) to assign a ranking, the TLF matches observed practice against a set of standards for effective teaching, in order to assign a rating. The theory of action is that teachers who are rated highly on the TLF framework will be associated with better student outcomes. However, if the ratings principals assign using the TLF don't accurately and consistently signal the quality of teacher practice, they will provide inconsistent or inaccurate feedback on whether teachers are effective and what they can do to improve. Principals are tasked with limiting judgments and inferences to those claims that can be substantiated with evidence. As Mr. Thom explained:

We're trying to write down everything the teacher says and all the responses of students, so we have more of evidence or proof of whether a teacher is doing a certain part of the standard or maybe needs some improvement in that area.

Every principal stressed his or her emphasis on capturing “evidence” before working toward any judgment, and explained that this emphasis on gathering such evidence was integral to using the Teaching and Learning Framework with fidelity.

Overall, principals in the study expressed strong support for the Teaching and Learning Framework as a measure of teacher effectiveness. Without exception, they
claimed their evaluation processes using the TLF were more objective, more comprehensive, and more specific than previous evaluation processes, which did not have clearly defined standards for practice. Almost all reported that the discipline of matching observed practice against a comprehensive set of descriptors that spanned multiple domains of teaching (a) challenged their preconceptions about effective teaching and (b) challenged them to be more rigorous about identifying components of teachers’ practice that were likely to lead to better outcomes, instead of relying on subjective impressions. In many cases, principals stated that using the TLF, they assigned lower ratings to teachers than they would have given under the previous evaluation system, including to veteran teachers. Of note, while principals discussed the AGT purely as a measure of effectiveness, they almost invariably linked the purposes of evaluation and professional growth and improvement when discussing the TLF. In fact, it is difficult to discuss what principals think about the TLF measures in isolation from how they use it to improve instruction, because the TLF itself evidently shaped the instruction it measured. Operationally, it was a formative tool, even as it captured evidence of performance.

In this chapter, I examine what principals thought observations based on the Teaching and Learning Frameworks actually measured.

The Teaching and Learning Framework as a Measure

While they understood the AGT as an outcome measure, principals in the sample universally described the TLF as a measure of teacher practices. In general, principals said the TLF describes what teaching practice looks like at different levels of effectiveness, and explained that the TLF could be used to measure consistently whether
the practices and behaviors of teachers constituted quality teaching. For example, Ms. Brown described the TLF as “a guide of what instruction should look like in a classroom.” Similarly, Ms. Alana stated: “I feel like it’s measuring effective practice . . . in as objective a way as you can. . . . It tries to give us an idea of what is effective teaching in a general context.” Principals also noted that the framework was essentially a narrative description of what teaching looked like at different levels of teaching skill, which made it easier for them to effectively distinguish between levels of performance.

Measuring practice against the framework and including indicators for planning and preparation not only changed how principals measured teaching, but also changed how teachers prepared and taught. Because teachers knew principals would be looking for evidence of certain practices, principals explained, they became attentive to demonstrating these practices in ways that could be observed or captured as evidence. For example, when teachers knew principals would be looking for evidence of specific aspects of planning, principals reported that they attended to those aspects of planning.

Similarly, when teachers knew principals would be looking for evidence not just that some students engaged in lessons, but that all students were meaningfully engaged in one way or another, they approached engagement more broadly and inclusively in the planning and delivery of lessons. Teachers were familiar with the rubric prior to the observations, and principals noted that they planned their instruction with the indicators of the framework in mind. As a result, the framework not only measured teachers’ demonstration of a specific set of practices, but also directed them toward use of those practices. In effect, they used the framework to plan their lesson, then the observation measured the enacted lesson against the framework.
Ms. Rosa’s comments were typical, when she noted that the TLF changed what she and teachers thought of as evidence, and this in turn changed how teachers prepared for observations. For example, during pre-observation conferences, Ms. Rosa explained:

I always ask them to bring in students’ work. I would ask them to bring in student work samples of whatever lesson I observed. [When I started using the TLF], they brought the student work samples and they brought some data to support certain students’ lingering needs or improvement that they had seen as a result of not just the lesson, but the instruction that had taken place since our last conversation.

The teachers used this evidence to plan and develop their lessons to address skills they had not demonstrated in prior lessons (e.g., the students were too quiet and needed opportunities to talk). In other words, the relationship between teaching and assessment of teaching was reciprocal.

As explained in Chapter 5, principals saw the AGT primarily as a tool for measuring growth in student learning, but less as a tool for supporting greater growth. In contrast, most principals consistently spoke of the TLF as both a tool for measuring effective teaching and a tool for supporting improvements in teaching practice that lead to greater growth in learning. With the TLF, most principals talked about formative purposes hand in hand with their use of the tool. Ms. Suss explicitly wondered if the TLF was really a formative assessment, because of its power as a diagnostic tool to assess which dimensions of a teacher’s practice were strong and which could be tuned to improve outcomes. As she explained:

When you’re asking questions, when you’re monitoring the room, are you really paying attention to what kids are doing? Are you critically thinking about what
they’re demonstrating? Are you stopping the lesson and refocusing if they’re not?

Are you factoring that in and stopping and doing something different?

When a specific practice was not observed, that absence raised the question of “why not?”

Like Ms. Suss, most principals could not describe evidence without immediately referring to its implications for improving practice.

For many principals, having the TLF changed their own understanding of effective teaching, and many expressed the conviction that being trained to use the TLF improved their own ability to observe instruction and provide actionable feedback. As Ms. Island reflected: “I think it has changed my understanding a lot in looking at what really best practices are, what a good teacher does, how a good teacher does it.” Because the rubric called observers’ attention to elements of practice they might not have considered in previous, less structured observations, these principals explained that the TLF not only pushed their own thinking with respect to good teaching, but challenged them to be more systematic and comprehensive in their observations.

In their discussion of the TLF rubric and what it measured, some principals explored the tension the rubric created between diversity of practice and consistency of evaluations. Using the rubric challenged them to grapple with the reality that good teaching is situational: what works with one group may not work with the next, and the right instructional strategy for one concept might not be appropriate for another concept or skill. Even though principals valued the TLF for providing a broad description of what they viewed as a very complex activity, some struggled with the specificity of the rubric and how to operationalize it across the diverse set of lessons they observed. On the one hand, they realized that not all lessons and classes and teachers were the same. On the
other hand, they knew that fair evaluations had to be consistent in the standard to which they held teachers. What these conversations had in common was that the TLF pushed principals to talk about “good teaching” in much more complex and nuanced ways.

**Broadening of Principals’ Operational Understanding of Effective Teaching**

Most principals viewed the TLF as a departure from previous observation tools in that it emphasized dimensions of teaching on which they might not have focused in previous evaluation models. Specifically, several principals noted that the TLF challenged them to look beyond specific teacher behaviors to dimensions of the physical environment in which learning takes place and to consider evidence of the impact of teacher behaviors on students to come to a judgment about teacher effectiveness. In addition, several principals noted that the TLF explicitly challenged them to pay attention and look for evidence of student engagement and student ownership of learning.

Many principals appreciated the shift in the TLF to a constructivist model of teaching, one that emphasized not just teacher moves but student behaviors and engagement in response to teacher behaviors. Ms. Suss’ comments were typical when she explained:

The TLF measures characteristics and qualities. Not only is it about seeing the teacher as a facilitator and how he or she interacts with the students, it really is about what the students are doing. But that all has to do with how a teacher moves through their classroom. . . . The TL framework measures [teachers] in terms of how students are taking ownership of their learning, like the room environment, learning strategies, how lessons are introduced, teacher taught versus student
taught. We are looking at a lot of different factors in terms of what’s making these successful lessons.

Similarly, Mr. Grace, who was a principal in one of the most challenging schools in the sample, appreciated the way the TLF challenged him to gather data on how engaged each individual child was:

The structure of the Teaching and Learning Framework is such that you can no longer ignore any student in the classroom. Because you’re going to look at data, you’re going to look at your grouping of students, you’re going to look at “what do I need to do?” It builds in differentiation of instruction.

Whereas previously, he had focused on the actions of the teacher, Mr. Grace found that in contrast, capturing observational evidence on the impact of instructional choices on students gave him greater insight into the fit between practices, instructional choices, and the specific needs of different students.

Ms. Vaughn noted the emphasis in the TLF on looking closely at the classroom as a learning environment: “The environment should be a learning tool, and it’s going to be a learning tool.” Implicit in this statement is the assumption that how teachers organize classrooms, use space, and decorate has implications for the kinds of learning interactions that take place.

Because the TLF emphasizes a common set of preferred practices, many principals, particularly in middle schools where subjects are departmentalized, felt having the tool gave them more focus and authority in their feedback to their teachers, and thus enhanced their ability to motivate change. Ms. Irma, for example, felt strongly that using the TLF gave her greater authority with respect to evaluating and providing feedback to
teachers who taught subjects other than the one in which she herself had specialized.

Ms. Irma noted that previously, teachers might challenge her authority to evaluate their practice, because her own training and preparation was not in their discipline. She said teachers used to say to her: “We understand the language arts part, since you have a reading specialist credential, but who are you to tell a math or a science or a PE teacher whether or not they’re doing an effective job?” However, the TLF focused on common elements of effective teaching, and thus gave principals like Ms. Irma more leverage and authority in conversations about practice. She explained:

In the case of the science teacher, I could say, “Well, I didn’t see productive work groups. I saw kids sitting at tables together, but I did not see a workgroup. You may have sat them together, and that may be your definition of productive workgroup. But here the language states this.” So I think it’s giving us a little bit more of validity of how it is that we can go ahead and rank somebody if you want to call it that, or score somebody.

Like a number of principals in the sample, Ms. Irma found that when she used the Teaching and Learning Framework, the conversation was focused more on the indicators in the framework and whether elements of effective practice were demonstrated and observable during the lesson, and less on her own credibility and authority. Deep teacher knowledge of the subject and pedagogical content knowledge are still important and associated with better outcomes for students. However, in the absence of the subject-specific knowledge to assess that, principals could still use the TLF to provide teachers with specific, actionable feedback on practice. This study sample only included principals at the elementary and middle school levels, and it is possible that the TLF is less effective
at higher-grade levels, where pedagogical content knowledge is more specialized and complex.

Factors that Affect the Quality and Accuracy of TLF Ratings

Although a few principals offered qualified critiques of the TLF, principals were positive in describing the Teaching and Learning Framework as an evidence-based approach to rating the quality of teaching practice and instruction, and the majority specifically mentioned the TLF as a substantially better and more objective process than the previous system of observation-based evaluations in the district. In general, principals expressed more confidence in the ratings they assigned using the TLF than the rankings assigned by the AGT. However, every principal also cited factors that might erode the accuracy or reliability of ratings of teacher effectiveness assigned using the TLF. Some believed their own preconceptions might shape, limit, or bias what they observed. Others wondered if their observations of different teachers might be inconsistent because of differences in the contextual information they brought to different observations. Because in the pilot they were using the TLF with only a handful of volunteers, some principals wondered whether they would be able to sustain the same quality and fidelity of implementation when using the TLF for every observation, due to the time-consuming nature of the work. Finally, as calibration sessions became more removed in time, many principals were less confident that the ratings they assigned were still consistent with the standards set during initial training. In the remainder of this chapter, I explore these four phenomena.

**Bias due to preconceptions.** As noted in the literature review, a longstanding complaint about evaluations based on principal observations of teachers is that
historically, judgments based on observations tended to be subjective, and for the most part, didn’t differentiate among teachers with different levels of skill. Most principals described the previous observation model used in the district—the Stull system—as subjective and lacking clear standards. Ms. Alana’s comments were typical of others:

The old Stull process really relied upon your particular principal, and what their goals were, what they were looking for when they came in the classroom, and you sort of had a general idea of what you thought they thought good teaching was, and what they were going to like when they came in to “Stull you.”

In contrast, most principals felt the TLF made their observations more objective and less arbitrary. Educators knew in advance what aspects of effective practice they were expected to demonstrate. Principals reported that the emphasis in the TLF process on scripting and gathering evidence for each indicator forced them to step out of making subjective judgments and into collecting observable data that would enable them to match the teaching they observed to descriptions of practice. If a particular practice was not evident during the observation, the TLF process discouraged principals from inferring that it would be present in another lesson. Several principals noted that with the prior, more subjective observation and evaluation process, they were more likely to let preconceptions about a teacher shape the evaluation. For example, Ms. Rosa noted that prior to calibration, based on her knowledge of a teacher, she might have given the teacher the benefit of the doubt if a specific aspect of good teaching was not evident on her visit. With the TLF, however, she no longer did so. If, over multiple observations, she did not observe evidence related to a particular indicator in the TLF, she now concluded that the teacher had not mastered that indicator. Without evidence, she could not infer
effectiveness. Ms. Rosa found that when she used the TLF, she gave lower ratings to some teachers than she had in the past.

Like Ms. Rosa, most principals indicated that the evaluations and ratings they were doing with the TLF were actually more rigorous and demanding than those she had given previously, sometimes in ways she found a little uncomfortable. Ms. Charles felt that the TLF framework challenged her to be more systematic and rigorous in her assessment of the nature of the interactions between her teachers and their students, and as a result, she felt her ratings of teachers were more consistent, and more accurate with respect to the true instructional quality of the teachers. She described a teacher who she had always rated favorably in the past, because whenever she entered the classroom, the kids were engaged and talking. However, the framework directed her to be more specific about the actual substance of those interactions, and to her surprise, she found herself reassessing the quality of that teacher and others for whom she had previously given quite favorable ratings. For example, she explained:

When I have this framework in front of me and it says, “Well, did [the teacher] ask challenging questions?” Well, I never noticed that she doesn’t really ask challenging questions. Yes, the kids, they’re all talking, but where’s the challenge? It gave me a lens to look through that I didn’t have before.

Ms. Charles found that as a result, this teacher, like some others, “failed” more often, and received more feedback that some practices were not effective. She lamented: “Even scoring was hard because I almost felt like I was beating her up. It was hard to say to a teacher that’s taught for 26 years.” However, using the rubric directed her to provide much more nuanced and specific feedback that shed light on instructional weaknesses.
that she would previously have overlooked. In this respect, she argued, the TLF was leading not only to much more accurate ratings of teacher effectiveness, but also to much tougher evaluations. Interestingly, many principals described giving ratings while using the TLF that were lower than they would likely have given previously.

Experienced teachers whom principals had thought of as skilled often did not exhibit behaviors associated with higher levels of performance on the TLF rubric, so were rated as “basic.”

While previous observation models felt arbitrary to these principals and difficult to defend, with the TLF, the evidence itself established the rating. According to several principals, when they provided the evidence they had collected against the indicators, the teachers saw the patterns and recognized on their own where their performance fell on the rubric, even if it was below what they had expected. Several described sharing the observation data with teachers before presenting their ratings, and noted that when they subsequently presented their ratings, teachers were not surprised. Some said that because the evidence, coupled with the rubric, took some of the burden of judgment off the principal, it freed principals to have more “collaborative” and productive conversations about how to improve practice.

**Bias due to differences in the contextual information.** While principals stated that the emphasis of the TLF process on evidence and objectivity was real and improved the rigor and accuracy of their ratings, they were mindful that prior knowledge also could be valuable and could actually enhance the quality of their observations. Their comments suggest that context might matter. Ms. Alana felt that in some ways, her prior knowledge might actually add nuance that enhanced the accuracy of observations, because it might
give her insight into why a teacher made the specific choices she made in the moment. However, to the extent the prior knowledge led her to a different inference than might be made by another evaluator, it eroded the reliability of her ratings, even as, paradoxically, it potentially increased the accuracy and local utility of the rating to the teacher being observed.

Even though they felt the TLF was a solid description of effective practice, principals like Ms. Alana sometimes struggled because they felt teaching itself was a nuanced activity and context specific, and that what was “effective” with one student, might not be “effective” with a different student. They worried that the framework might not adequately capture this flexibility on the part of the teacher. Ms. Alana explained that from her perspective, the TLF described “effective practice” as objectively as possible. However, she went on to note:

Sometimes a teacher may do or say something to an individual student or to—
with a group of students that is absolutely effective in that context at that moment.
If you took it out and did that exact same thing with another group, you might say that that’s not working.

For Ms. Alana, the art of teaching was that ability to diagnose and personalize instruction in the moment, and she, like several principals, felt professional expertise and knowledge played a powerful role in interpreting what they observed.

Principals also were emphatic that school context matters. For example, Ms. Island, the principal of one of the higher-poverty and lower-performing schools in the sample, identified some of the TLF indicators related to parents’ participation. She noted that though these expectations are worthy, it can be more challenging to engage and
partner with parents in schools with high rates of student turnover, in part due to family instability. Ms. Island described looking at all her evaluations and realizing that she had no evidence to support a high rating for parent participation and engagement for any of her teachers: “No one has parent participation, so that’s going to be a problem with this framework when the teachers say, ‘I know I did a great lesson, but yet I’m only getting a three or a two for a rating.’” Her perception was that it would be much easier to engage parents in visible ways in a school with more affluent students, and that “good parent engagement” might look different in her context than in others.

Principals also felt that observing a teacher on multiple occasions, as they delivered multiple types of lessons, would enhance the reliability of their ratings. Several principals noted that some lessons fit the framework well, and some others—including very good lessons—did not fit the framework. For example, a lesson that was designed to develop deep conceptual understanding was more likely to call on the teacher to demonstrate questioning strategies, while a lesson designed to develop fluency might emphasize focused practice. The latter would not give the teacher a chance to demonstrate development of higher order thinking through questioning strategies. Both lessons, however, would serve an important purpose in the development of student capabilities, and good teachers would likely present both types of lessons over the course of a year. Thus many principals hypothesized that to accurately and reliably capture the true breadth and scope of an educator’s repertoire, they might need to observe a given teacher on multiple occasions over time. As explained elsewhere, most principals also noted that standards-based observations were time consuming, and they were unsure whether they would be able to complete them well after the pilot, when they would need
to evaluate much larger numbers of teachers.

   Principals were also aware that it would be difficult for a teacher to demonstrate all capacities assessed by the TLF adequately in a single lesson. Some also described working with teachers proactively as they scheduled observations, to make sure they demonstrated their full repertoire. For example, good instruction can involve both development of higher-order thinking and development of fluency in basic operations. Teachers need a flexible repertoire that may be difficult to see in a few observations using the TLF. To explain, Ms. Kim described her reaction as she toured her building:

   We went into first grade and they got a structured blending lesson, and then we went into the third grade and they had done book reports and they were eating cereal and going around and passing around cereal box book reports. Those are two different kinds of lessons, and I want to see both those lessons from a teacher. Ms. Kim, like other principals, felt when she could visit classrooms multiple times, she was more likely to capture evidence of the full breadth of a teacher’s repertoire.

   Most of the preceding factors addressed the capacity of the framework to capture complex practice. A small number of principals also raised questions about whether the changes LAUSD had made to the TLF framework would erode their capacity to use it in ways that would lead to improved outcomes. These principals noted that the full Danielson Framework had been validated, and in empirical research was positively correlated with better outcomes for students. However, for uses in LAUSD, the district had limited the number of high priority indicators in the TLF to reduce the time burden on principals. The two principals who expressed this concern worried that modifying the tool might attenuate the relationship between ratings on the Framework and improved
student outcomes. In the words of Mr. Isaac:

I would have preferred if we just dispensed with all this political stuff of “let’s get everybody’s input” and just stuck with the original [Charlotte Danielson] document, because no matter what instrument you come up with to describe something as complex as teaching practice it’s going to be limited anyways.

Mr. Isaac’s statement reflected his larger commitment to being as rigorous as possible about evidence-based practice, while acknowledging that the reality of implementation is often more complex.

**Bias associated with maintaining reliability and accuracy while scaling up implementation.** As part of the pilot, principals received intensive training and calibration in use of the rubric. Principals were trained to observe carefully and script what they saw and heard during their observation. Only then did they match the data they scripted against the descriptions of practice in the rubrics, which corresponded to different ratings of effectiveness. In the pilot year, they co-observed with expert raters, and calibrated the ratings they assigned with the ratings of these experts to promote consistency and accuracy of ratings. However, even with this support, principals anticipated challenges in ensuring accuracy and reliability of their ratings over time.

First, principals said the sheer breadth and comprehensiveness of the TLF, and the time involved in using it with fidelity, would challenge their ability to provide accurate and reliable ratings of teacher quality over time. Despite a few concerns about the streamlined Framework, as noted above, most principals were relieved the district had narrowed the number of indicators to use in LAUSD, because the amount of work associated with using the framework as intended was immense, and the time the process
took threatened principals’ ability to implement it well. Even with the streamlining, they described the process as challenging to implement. As Ms. Charles said: “I think the harder part of this process was the demand of just trying to put it all into the computer. It just took so much time and so much effort. It’s important, but I wonder how that can be streamlined more.”

Many principals, like Ms. Charles, discussed the enormous amount of time associated with using the TLF in a robust way to rate and provide feedback to teachers. Preconferences, planning, careful scripting, and meticulous matching of scripts to the framework all took a lot of time, and time is a scarce and precious resource in schools. All principals noted that during the pilot, they were conducting the TLF observations with few staff, and that if and when the new evaluation process became fully operational, they would be carrying out these valuable but time consuming processes with a much larger group of teachers. Most expressed some concern about whether they would have the time to implement the new process with fidelity when they had to work with a larger number of teachers.

In addition, both principals who were new to the TLF process and some who had several years of experience with the Danielson Framework and subsequently the TLF expressed some uncertainty as to whether their ratings would remain consistent over time. Several explained that when they were new to the standards-based observation process, they were not always clear about what to prioritize when they initially assigned ratings, and not always confident they were accurately capturing and tagging the right information. After two years in the pilot, principals felt more confident and expert with the process locally, but less sure their scores would still be calibrated with those of other
principals. As Mr. Isaac summarized: “To be honest with you, after a year of doing the pilot and working with six different teachers and doing it twice, it was a daunting task. I’m not confident in my own validity and reliability in terms of implementing, if that’s the right words.”

Regardless, even if their ratings were no longer calibrated, most principals felt the TLF tool and process still had value in strengthening their efforts to improve instruction. They described laboring over data entry and deliberating over evidence. They described investing heavily in using the tool appropriately and negotiating the online platform the district used to support the TLF. Most suggested they did so because they felt the tool changed the conversation from whether or not teachers were using effective practices, to a discussion of which practices teachers demonstrated, and what new practices might they develop to become more effective. In this context, many principals said the post-observation conversations, even if challenging, felt more collegial and more focused on the shared goal of improving instruction. The TLF enabled these principals to give specific feedback and recommendations using a common language or definition of effective teaching, and this clarity made feedback actionable. Most principals clearly felt that even if the tool posed some challenges to reliability and accuracy, especially as the initiative scaled up to include all teachers, there was substantial value in the formative potential of the TLF as a tool for supporting and promoting improvements in teaching and learning. For this reason, despite concerns, most principals described working hard to implement the TLF with fidelity.
Chapter 7: Improving Teacher Practice

Using the TLF to Support Growth and Improvement

As noted in the prior chapter, one of the most powerful findings of the study was that principals had difficulty discussing the TLF without emphasizing how they used it to improve teachers’ practice. A second powerful finding, hinted at in the last chapter, was that the tool seemed linked to a change in relationship between principals and teachers. Principals generally stated that, absent any standards-based criteria for evaluation, evaluative conversations with teachers often pivoted on teachers’ perceptions of the subjectivity of principals’ judgments. In contrast, the TLF process required principals to capture evidence of effectiveness through observation and scripting, then match the evidence to specific descriptors of practice in the TLF framework. A teacher’s rating against the TLF was a function of the corresponding evidence, not, in theory, the principal’s judgments about quality. Principals in the sample explained that this led to more instruction-focused conversations with teachers: they presented the evidence, and it was clear where teachers ranked. The conversation thus focused less on where teachers ranked, and more on how they could improve.

The TLF and Improving Practice: Shifting Purposes and Roles

While the AGT was sometimes used as a tool for motivating improvement efforts, the TLF was primarily a tool for talking about how to make those improvements. As noted in the previous chapter, several principals said that using the TLF, they had much “harder” discussions about teaching and assigned lower ratings than they did under the
previous evaluation model. In addition, the TLF process led principals to identify domains or aspects of practice on which teachers, including veteran and previously highly rated teachers, demonstrated substantial room to improve.

Ms. Richmond explained this as a tension between previous expectations and understandings of effectiveness, and a new set of expectations. The new tool—the Teaching and Learning Framework—exposed a gap between where teachers currently were in their practice, and where the district would like them to be, based on new expectations for effectiveness. She noted that in her own school, she had several National Board Certified Teachers who had long been viewed as strong teachers. However, “once you apply [their teaching] to a rubric, and it isn’t solely my own perception, they’re ‘novices’ and they’re ‘developing.’ And there’s a lot of growth that’s needed.” Ms. Richmond found the TLF compelled her to hold her teachers to a new and higher standard, and as a result, many teachers were getting lower ratings than they had come to expect. As Ms. Charles said:

It was kind of an eye opener for me … because there was no denying what was seen or not seen. And I think it’s kind of easy when it’s subjective to deny it because you are looking at a person that’s a nice person, but ineffective. What the rubric highlighted was the absence of evidence of effective learning practices, even in classes that appeared to run smoothly.

New Focus on Student Engagement and Student Response to Instruction

While previous evaluation systems emphasized teacher actions and behaviors, every principal in the sample noted that the TLF challenged them to call attention not just
to whether teachers were using specific best practices, but to what the actual, observed impact of those practices was on the students in their classrooms. In doing so, these principals all expressed the sentiment that the TLF pushed teachers past a focus on well-written lesson plans to a focus on the actual fit and effectiveness of those plans with real students in the moment of instruction. In interview after interview, the single most profound and universally expressed impact of the TLF was the way it redirected instructional improvement efforts to a focus on the impact of teacher moves and decisions on the engagement and learning of individual children.

Ms. Suss noted that she had always “scripted” lessons as she observed, but that the TLF directed her away from focusing only on what the teacher was doing and from looking at the lesson from the perspective of the teacher, and towards “looking at how the kids were interacting with it, because at the end of the day that’s what it’s about. It’s about their ownership and about whether they’re really learning it. You can give them all a lesson, but it doesn’t mean that everybody got it.” Ms. Suss stated that this shift forced her and her teachers to be more explicit about what successful engagement really looks like, and to ask hard questions about whether they were really redirecting students effectively and adjusting their teacher “moves” to successfully engage all learners in learning tasks. Like many principals, Ms. Suss emphasized that high-quality teaching is often situational: “What works one year doesn’t necessarily work the next year, depending on your children and what they’re capable of doing.” In Ms. Suss’s experience, using the TLF focused her conversations with teachers less on whether they had a good lesson plan, and more on the effectiveness of the implemented plan, in the moment, with each of the children in the room.
A common theme among principals was that the TLF also challenged teachers to move beyond thinking about “the students” as a group, and towards specific thinking and strategies targeted to individuals within the group. It demanded attention to differentiation. As Ms. Suss said: “You want to be rigorous, but you always have to think about the three or four kids. Now this is going to be a problem. How do I put something in my lesson that still will help everyone, but is targeted for all these kids?”

Principals saw this shift to focusing on the actual impact and effectiveness of the instructional choices teachers made as pivotal. As Ms. Charles noted somewhat wryly:

I can teach to my heart’s content and be up there and be wonderful and then look at the kids and they can be going absolutely crazy on the rug. . . . Wonderful teaching going on, but is anything getting to the kids? To give true feedback to people, you have to talk about what kids are doing. That’s your true evidence.

This sentiment, and the power principals attributed to the TLF as a tool for focusing on teacher impact, was striking for the consistency and clarity with which it was articulated by principals across the sample. According to principals, this emphasis on students’ response to instruction fundamentally changes the game for teachers, by reinforcing the fact that to get different educational results, teachers need to adjust their instructional choices.

The profound shift caused by principals’ use of the TLF was summed up by Mr. Grace, who said:

If learning doesn’t take place in the end, then it doesn’t really matter how good a teacher you are in terms of what you did. That is what the Teaching and Learning Framework balances: both the action of the teacher and what the students are
In Mr. Grace’s school, using the TLF promoted a tighter linking of specific practices to observed student engagement, and challenged teachers to make different instructional choices in order to obtain different educational outcomes.

Principals in low-performing schools spoke about using the AGT to motivate teachers to take responsibility for the learning of individual children. However, with the TLF, principals said they could directly connect teacher behaviors to observed student responses. Ms. Charles noted that teachers could identify reasons to discount the AGT rankings, but that when they read the TLF scripts, they saw evidence that challenged them to reflect on their own teaching: “As they read the script they were like, ‘Wow!’ or ‘They said that?’” Ms. Charles observed that during the moment of teaching, many teachers didn’t hear or focus on some things that their kids were saying or doing, but the scripted evidence challenged them to rethink their actual lessons, as opposed to their intended lessons. The immediacy and comprehensiveness of the evidence were unavoidable.

**Actionable Feedback Enhances Effectiveness**

Principals explained that the emphasis on scripting and then matching scripted evidence to the descriptions in the TLF enhanced the impact and credibility of any feedback. For example, Mr. Roland explained:

I asked a teacher, “How many kids do you think were paying attention when you were doing your introduction lesson?” And he would say, “Most of them.” And in the past, it would have been like, “No, I don't think all of them.” But when you
can give them numbers, like, “Out of 24 kids, three kids were looking at you, six kids were playing in their desk, five kids left the room.” And give him some numbers like that. That gives them something else to think about in a different way than just, “Good job. Bad job.”

Similarly, Mr. Grace found that when he shared scripted evidence of how students responded or did not respond to specific instructional actions, he was able to challenge teachers to be reflective, rather than defensive. In his conversations with teachers, Mr. Grace said he and the teachers would review transcripts and reflect on each question and response, and analyze the transcript for evidence of impact, and then consider the indicators in the framework to identify additional strategies to try to enhance any observed impact. What principals like Mr. Grace appear to be doing is using the TLF scripts to explicitly bridge the gap between teachers’ theories of action and intended instruction on the one hand, and their observed enacted practice on the other.

Many principals described using the descriptions of practice in the TLF to promote critical reflection by teachers on what constitutes effective practice. Mr. Isaac used the indicators that related to student engagement and participation to challenge teachers to think about what it looked like to purposefully and meaningfully engage their students, even as he acknowledged that there might be no one way to adequately describe engagement. The power of the tool was in the process: teachers had to grapple with the actual impact of their instructional choices, and what practice might look like if it were more effective. As Mr. Isaac explained: “There is no perfect articulation of what the most effective practice is. It’s a matter of opinion and there’s a lot of variables there. But the TLF is useful if it’s something that prompts teachers’ own reflection.” Framing the use of
the indicators this way leaves teachers with responsibility for making professional judgments about how to next improve their practice, but it also makes clear that there are further opportunities for teachers to become more effective. Implicit also is Mr. Isaac’s respect for the professionalism of his teachers, even as he encourages them to use a standardized tool as the starting point for a conversation.

Effectively, the debriefs of observations using the TLF transformed post-observation meetings into opportunities for professional development. Ms. Suss noted that the extensive scripted evidence supported more nuanced and thoughtful conversations with teachers. In previous observation processes, she reflected, it was easier to become complacent with what they “knew” and what they “assumed,” and easy to rely on a few teacher moves, rather than employing a multitude of strategies to better engage diverse learners. As an experienced principal, she suggested that with the TLF, as teachers and principals discussed the indicators and the related evidence, they were challenged to deepen and broaden their understanding of critical aspects of practice, including student engagement and what it looks like, or differentiating instruction to meet individual needs. She noted:

There are parts in the framework where it talks about whether students are engaged. I was able to talk to teachers and say, “Okay, so here’s how you introduce a topic. How did you feel about how you got the kids to interact in groups and what were your different ways to do it?” . . . And we realized that we really sometimes don’t give children a variety of strategies to interact, because there are just lots of ways to do that. [The TLF] gave us better conversation about what it looks like [to engage students], other than a more effective question.
In the course of these conversations, Ms. Suss said that her teachers were able to broaden their repertoire of strategies to engage students.

**Managing the Tension Between Evaluation and Continuous Growth and Learning**

An undercurrent in many contexts was the tension associated with uses of the tool and, in particular, the tension between evaluation for the purpose of “rating” and evaluation for the purpose of fostering learning. Principals positioned the TLF as a tool for fostering formative conversations, and worked hard to use the tool to promote continuous improvement of instruction. However, once the tool was used in summative ways, the use shaped the experience of the tool. On the one hand, clear expectations as outlined in the TLF rubric created some positive pressure for teachers to step up their game. On the other hand, ratings could make teachers defensive or demoralized, especially when the rating was lower than expected.

Ms. Ivy was acutely aware of navigating this tension as she led her teachers in efforts to improve questioning techniques. Her faculty had been using Lauren Resnick’s *Principles of Learning* work for some time, to develop questioning techniques. When she introduced the TLF into this work, it served as a guide and signal for how, from an evaluative perspective, use of questioning would be operationalized. On the one hand, Ms. Ivy felt this was good. It emphasized effective use of questioning as an expectation, and when use of questioning was evaluated against a rubric, teachers tended to zero in on specific behaviors that would make evident their use of effective questioning. On the other hand, Ms. Ivy noted that when questioning was discussed in the lower-stakes context of professional development, conversations were generative and led to a broader
and more responsive repertoire of strategies to enhance instruction. Ms. Ivy did not want to lose those generative conversations or incentivize her teachers to focus simply on conforming their practices to the rubric, at the expense of a culture focused on developing professional practice. To help keep the focus on continuous improvement, Ms. Ivy talked about the TLF as a formative tool, and not a summative “one shot deal” tool, and she emphasized seeing individual growth and improvement as a long-term goal.

While acknowledging the need to rate teachers, Mr. Grace was reluctant to overemphasize teacher ratings, in part because of the tension between evaluation and learning. In his view, teachers have a professional responsibility to continuously evaluate evidence of impact and use it to improve their work, regardless of how troubling the evidence or how great the challenge of change is. For Mr. Grace, the challenge was getting teachers to be open to feedback and continuous improvement, as opposed to being focused on and afraid of evaluation. As he explained:

People are always focusing . . . “This is how I’m going to be evaluated.” They’re missing the point. You shouldn’t worry about how you’re going to be evaluated if you know that in the end you’re going to do the good thing for the kids.

Mr. Grace knew his school was very low scoring, and in his opinion, focusing on low scores or low ratings alone discouraged teachers, rather than motivating them.

**Emphasis on Developing the Professional Capital of Teachers**

One might wonder whether, if presented with data that suggested certain of their instructional choices were ineffective, teachers might shut down or resist. Several principals suggested that, although that was a possibility, it depended in large part on how
the principal handled and used the data.

Mr. Grace, a principal in one of the lowest-performing and highest-poverty schools, stated that his teachers’ professional commitment to the learning and well being of their students predisposed them to listen and reflect. He explained:

If the discussion is centered around the students, and your teachers understand that, then it is really about the students and the learning. I don’t doubt that the teachers got into this business because they want to influence and have impact on student learning.

Mr. Grace described working to make clear to teachers that his priority was to improve learning and that his intent toward them was supportive and not punitive. He stated that he did not have challenges working with the data, even if it was critical feedback, because his teachers understood the goal was improvement, not punishment.

At core, what Mr. Grace seemed to be suggesting is that the TLF process gave teachers powerful tools and evidence to improve their practice, and that in itself was professionally empowering. Mr. Grace explained that he felt one positive benefit of the TLF and the associated professional conversations was the positive impact on teachers’ sense of their professional capability, and in particular, their confidence in their ability to improve and adjust their professional practice in ways that lead to higher student engagement and better outcomes. Mr. Grace appeared to be trying to use the TLF to build teachers’ sense of efficacy, which he defined as “their own ability to influence students.” Mr. Grace expressed confidence that continuously using the evidence gathered through use of the TLF would help strengthen his teachers, strengthen his school community, and support and strengthen partnerships with parents. Mr. Grace was not concerned with
whether the TLF could improve learning in his school; he assumed it could. However, he was concerned about whether he would have the time and support to implement use of the TLF with all the teachers he supervises with fidelity, such that the potential and power of the process could be fully realized. This challenge was all the more acute in his high-poverty, high-turnover context, where even during the interview, he was interrupted multiple times to respond to crises.

For Ms. Alana, a principal in one of the highest-performing and most advantaged schools in the sample, the challenge was a little different, even though she was similarly focused on building teacher professionalism. She also noted that providing teachers with hard or disappointing feedback could be threatening. However, she suggested that how feedback was received and used was in part a function of the relationship principals had with teachers and whether that relationship was supportive. In addition, she noted that the value of the TLF for developing teacher capability was dependent on each individual teacher’s receptivity to feedback. She explained:

Is that teacher open to change, open to feedback? And if not, it really doesn’t matter how involved a process you make. It’s just—the process makes it more formal that now I can tag this standard, this standard, this standard, this standard, and all these things, and here’s the script evidence. I guess you could say that that’s more effective. If it doesn’t improve the teacher, I suppose it provides documentation for whatever discipline might be coming down the road.

Implicitly, Ms. Alana is suggesting that if teachers are not open to change and improvement, no tool will make them so. When a teacher is not receptive to feedback and improvement, the TLF instead serves as a tool for moving them out of the classroom.
Mr. Isaac, another principal in a high-performing school, similarly emphasized the professional capabilities, motivations, and qualities of his teachers. He described the value in his role of sustaining a focus on results, but not in micromanaging the means to achieve those results. He explained:

The thing is, if you have those motivated people, you have to get them to buy in and agree on what the outcome is, what’s the problem that we’re trying to solve here and the problem we’re trying to solve is raising student achievement. And if they agree on that, then you have to let them—as long as they’re legal and ethical in how they do it—let them solve the problem. I’ve come to believe and to understand for this type of community of teachers, it’s about being results-focused.

He went on to caution, however, that being results-focused did not mean going into classrooms and telling teachers how to change their practice, because that kind of micromanagement, in his opinion, worked against developing the kinds of professional skill and judgment teachers need to respond flexibly and in powerful ways to engage individual students and promote learning.

Mr. Isaac’s comment suggests a belief that schools get better not because principals tell teachers how they are doing, but because principals build the capacity of teachers to evaluate evidence and act on it collaboratively in ways that lead to improved learning. As Mr. Isaac stated:

I think the real model if you have those high-achieving people is saying this is what we’ve got to do. You’re in there every day. You know what’s going on. You’re best equipped to come up with a solution of how to get there. And then have lots of dialogue along the way with them. For me, it’s just asking the next
question that they may not be considering, as it relates to their plan for intervention or improvement.

Viewed from this lens, the principal’s role appears to shift from rating and prescribing to sustaining a tight focus on mission and questioning and challenging, with the goal of developing the professional capability of his or her teachers.
Chapter 8: The Relationship between the AGT and the TLF

Making Sense of Discrepant Cases

All principals in the pilot received AGT rankings for the teachers they were observing using the TLF. They thus had the opportunity to examine a quantitative outcome measure of relative teacher effectiveness (the AGT), as well as a qualitative process (and more formative) measure of teacher effectiveness (the TLF). In this study, I was interested in the extent to which principals perceived that AGT rankings and the ratings they assigned using the TLF told consistent stories about teacher quality. In other words, were teachers who received higher ratings on the TLF relative to their peers also teachers who were ranked higher based on the AGT? Did teachers who principals rated as novices also have low associated AGT rankings? In contrast, if principals and teachers struggled to reconcile AGT rankings with the ratings assigned using the TLF, or if the TLF ratings seemed on their face to be inconsistent with what AGT rankings suggested, this discrepancy would raise questions about the tools themselves, as well as how the tools might shape judgments about teacher efficacy.

In addition, although one is an outcome measure and one is a process measure, because both the AGT and the TLF sample aspects of teacher quality, we would expect them to be somewhat correlated. However, as noted in the introduction, empirical studies and the district’s own research suggest this correlation is modest. An analysis of discrepant cases, where principals struggled to reconcile the AGT rankings with their own ratings, sheds light on why this correlation is modest.
For context, it is important to note the AGT ranking is not an absolute measure of teacher quality, but rather a relative measure that compares teachers to each other. In contrast, the TLF rating is a rating that compares a teacher’s performance to a fixed set of standards. If teachers in LAUSD are all outstanding, many could have high TLF ratings and still have low AGT rankings. Inversely, if instruction in LAUSD is weak overall as measured on the TLF, some teachers might have low TLF ratings and still have a high AGT ranking. There are other circumstances that might also yield one of these “high/low” situations, in which the score on one measure differs from what one might expect on the other.

In this section, I explore the extent to which principals believed the two measures were correlated. In addition, I analyze a selection of discrepant cases to explore reasons why, from the perspective of principals, correlations might be modest.

I was curious about whether principals identified or observed “discrepant cases” where the level of quality suggested by one measure did not seem consistent with the level of quality suggested by the other. If so, I wanted to understand how they explained this discrepancy, and whether they thought discrepant cases had an impact on how they used these tools or inferences they made about certain teachers. In addition, I wanted to explore whether principals’ explanations for discrepancies could shed light on the attenuated relationship between AGT rankings and ratings of teachers based on the TLF.

I found that most principals in the sample identified cases in which they felt the AGT and the TLF seemed to be signaling different realities or levels of quality. In addition, how principals grappled with discrepancies was different in higher-poverty and
scoring and lower-poverty and -scoring schools. In the section that follows, I give examples of how principals explained and made sense of the discrepancy.

**Expectation of Correlation**

For the most part, principals assumed that ratings on the TLF rubric should be somewhat correlated with AGT rankings. This reflected a general assumption that if teachers taught well, their students would likely score better relative to children taught by less effective teachers. In addition, because many principals operationalized their understanding of the AGT as a measure of absolute—not relative—impact on learning, this assumption makes sense.

There was still considerable variability in the confidence principals had in the strength of the relationship between the two measures. Principals who appeared less confident in their own TLF ratings or more deferential to test scores tended to question the reliability and accuracy of their TLF ratings, while principals who were confident in their use of the TLF tended to raise some of the concerns cited in earlier chapters with respect to the AGT. In addition, principals in higher-scoring schools were more confident in questioning the AGT ranking than principals in lower-scoring schools. Perhaps their high scores insulated them in ways that allowed them to challenge test-based measures when lower-scoring schools could not.

Most principals mentioned that on average, while the AGT rankings seemed consistent with what they expected based on their classroom observations of practice, there were bound to be exceptions. As Mr. Isaac, a principal in a high-scoring school, explained:
I don’t know the math, if you can figure out what the noise is and all of that, but if all of these things [in the TLF rubric] were going on inside a classroom at the highly effective level—because it’s a pretty comprehensive description of practice—then probably there would on average, not every single student necessarily, but there would be student growth that would be captured in standardized testing at the end of the year.

“On average” was the key term: principals seemed to think that, while the two were likely to tell similar stories overall, exceptions were bound to arise.

Several principals mentioned that both measures were themselves imperfect, and that this could also contribute to an attenuated relationship between the two. For example, several noted that not all the learning captured in the Danielson Framework is captured in the abbreviated Teaching and Learning Framework, and these principals wondered if the changes to the TLF might weaken the relationship to outcomes. Additionally, not all the learning captured in the Danielson Framework is sampled by the test. Some principals noted that some standards, including standards that link to inquiry, might be underemphasized on the CSTs, because they are difficult to sample on a standardized test. Thus, for example, if a teacher did not teach in ways that developed scientific inquiry, this gap might not be reflected in test scores. However, classroom observations would sample the presence of teaching that supported inquiry, and inquiry-based practices and higher levels of engagement associated with inquiry practices might yield higher TLF ratings. As Mr. Isaac summarized: “Not all that learning is captured on the test because the test may test some subset of what principals are looking for [in observations].”
For these reasons, several principals were cautiously skeptical about the relationship between AGT rankings and ratings based on the TLF, and suggested that differences between what is sampled by the TLF and the AGT could explain the low correlations, or yield high correlations even when they were, in the principal’s perception, not warranted. A few principals, and especially principals in higher-scoring schools, reiterated that the practices in the TLF might lead to greater engagement and more complex thinking in the classroom. However, they also felt some of the complex and engaging learning they value and the framework supports is not adequately sampled by the test.

Ms. Alana, a principal in a high-scoring school, identified discrepancies as an example of why principals need to rely on multiple types of data and mediate inferences about quality based on AGT rankings. Because her students were so high scoring—many had perfect scores on state tests used for accountability purposes—it was impossible for many of her students to demonstrate gains, so at best her teachers would be ranked as average. This led her to discount AGT rankings and to expect a more attenuated relationship between the AGT rankings and her TLF ratings.

In contrast, Ms. Mitchell, a new principal in a low-scoring school, relied heavily on student test scores and teacher AGT rankings to manage performance and set improvement goals. She used the two measures iteratively to raise scores, and relied heavily on interim benchmark testing to make sure students were progressing steadily toward goals. The effect was to tightly link management of instruction to the scores, including the interim benchmark scores, which in turn, might increase the correlation between ratings and scores. Ms. Mitchell explained:

We use periodic assessments as a baseline for what predictors we can have for
students achieving on the CST. For an effective teacher, you would need to be somewhere between 85 to 95 percent in the proficient and advanced areas, because that would demonstrate that the concept was known to the majority of your class. That’s what we’re looking for and expecting. If they were at 60 percent, we’d be looking at what differentiation and what reteaching opportunities they were going to provide for those students in order to receive quality instruction.

Note here that Ms. Mitchell is relying on proficiency scores to decide if a teacher is “effective.” Regardless of where students started the year academically, Ms. Mitchell appears to expect that if the teacher is effective, between 85% and 90% of their students will be scoring as proficient after instruction. For teachers in her high-poverty, low-performing context to demonstrate this kind of proficiency, they would have to generate substantial gains. If they did achieve this feat, they would likely have higher AGT rankings, and Ms. Mitchell was confident that these teachers would also have high TLF ratings. What is unclear in this context is whether the reverse is true: would all effective teachers as measured on the TLF be able to generate the level of gains implied by moving a low-performing population to 90% proficiency in one year?

Ms. Vaughn, another principal in a low-scoring school, similarly expected a high correlation between the AGT rankings and her TLF ratings. She observed: “I looked at AGT after I’d been here about three months, and I was so surprised that it absolutely matched what my perception had been.” She expressed strong confidence that the AGT rankings reflected teachers’ use of the strategies in the TLF, and that teachers who adjusted their practice as guided by the TLF would have students who demonstrated
greater gains.

Ms. Alana and Mr. Isaac’s skepticism on the one hand, and Ms. Vaughn and Ms. Mitchell’s tight focus on scores on the other, represent the variability of principals’ responses to the relationship between these measures. In places with greater skepticism, principals tended to rely on a broader set of measures and target a broader set of goals. These also tended to be higher-scoring schools. In schools where educators trusted or deferred to scores, the TLF and the AGT, along with interim benchmark assessments, worked hand in hand as a tighter system of feedback focused on improving outcomes as measured on tests. These tended to be lower-scoring schools. Higher-scoring schools tended to have more affluent students and fewer Latino and African-American students. These findings suggest that accountability may have different effects on professionalism and on how and what teachers teach in high-poverty and lower-poverty contexts, and in higher-scoring and lower-scoring contexts.

That said, many principals identified cases in which principals were surprised by AGT rankings that they felt were inconsistent with what they observed using the Teaching and Learning Framework. These cases tested principal confidence in the correlation between the measures. As Ms. Dale mentioned:

Somebody that I thought would have a low AGT, didn’t. I would say he was a developing teacher. So I’m sitting there going, “Okay. I’m not seeing the correlation.”

Depending on whether principals were more confident in the test as a measure of effectiveness or more confident in their own professional judgment, they responded differently to these “discrepant cases.” In the section that follows, I discuss how
principals made sense of discrepant cases.

Discrepant Cases in Which the Principal Privileged Professional Judgment over Test-Based Measures

In some schools, when confronted by discrepant data, principals responded by relying on their professional judgment and reviewing multiple sources of data beyond the test-based measures. These included cases where the AGT signaled both that a teacher was better and other cases where a teacher was worse than the principal expected, based on observations. For example, Ms. Irma identified a much admired math teacher in her building who received a very low AGT ranking. This math teacher had consistently been rated very highly in observations, and was also viewed in the building as an instructional leader and expert. In the school, she led professional development and provided training to other teachers. As Ms. Irma explained, the AGT ranking surprised not only her but her colleagues, because it ranked this teacher as less effective. As a result of the ranking, this teacher felt she needed to step down from leadership within the school, but Ms. Irma said: “We told her the AGT is a measure. We need to look into it deeper, but we don’t think it was enough for her to step down. I still believe in her as a teacher.” In this case, despite the low AGT ranking, Ms. Irma’s default assumption was to trust local evidence and observations in forming her assessment of the teacher, and to discount the AGT ranking when it produced results that were inconsistent with other data sources. What was disconcerting to Ms. Irma was the lack of clarity about how to reconcile these results, and what questions this raised about how to best understand efficacy.

Ms. Irma was also confronted by a second case, in which a teacher had a very
high AGT ranking, but needed extensive coaching and intervention on his classroom management and treatment of children. As in the previous case, she discounted the test-based ranking because it contradicted what she observed in terms of management on a daily basis. Instead, her assessment of a teacher’s efficacy was informed by multiple data points and grounded in her own professional judgment.

Like Ms. Irma, Mr. Roland tended to privilege his own judgment, based on his local review of data, over the AGT ranking. And like Ms. Irma, his primary concerns related to issues of classroom management. He said:

I don't really focus on AGT scores. I did look at a couple of them, and there was one that I was really surprised where it showed that he was pretty effective with kids, like he got a four I believe, or something like that. And yet, his class does not have—it always has discipline issues. Maybe I shouldn’t say always. I should say eight out of ten days he has discipline issues.

When this teacher cited the high AGT ranking as evidence he was “doing great,” Mr. Roland continued to focus on classroom management issues that he felt were a priority, regardless of the AGT ranking. The comments of Ms. Irma and Mr. Roland raise the possibility that principals may place a priority on effective classroom management that is not reflected in scores.

Similarly, Mr. Isaac also observed a pattern in which teachers with very poor practice or harsh teaching styles and poor classroom management often had higher AGT rankings than he would have expected. He described some of these teachers as the ones he was “just waiting for . . . to hang it up and leave,” because on his visits to their classrooms, “there was no evidence of learning, the room was a dump,” and instruction
took the form of what Mr. Isaac described as “a rant while students sat passively at their seats.” Mr. Isaac observed that in his school, “those teachers who are particularly harsh on the kids and very demanding and not very patient and stuff, a lot of those teachers came back with very high value-added scores.” Mr. Isaac commented that after placing students and working with parents, he knew parents would rather take a chance on a teacher with the lower AGT ranking than have their child placed into the classes of some of these (harsher) teachers. Whatever the AGT ranking signaled, he found that most parents preferred to not have their children in these classrooms.

The common theme in many of the discrepant cases was a teacher with high AGT rankings but poor classroom management or poor relationships with students. This raises questions as to whether some principals place a high priority on classroom management and teacher–student interactions that is captured in lower TLF rankings, but which is somewhat orthogonal to performance on tests, and perhaps not captured in the AGT rankings. It was very clear, however, that these principals viewed poor relationships with students and poor classroom management as unacceptable, and therefore they discounted a positive AGT ranking when a healthy classroom climate was not evident.

Some principals noted that test scores might not be reliable for the purposes of evaluating teachers, because especially at the elementary level, where rankings were typically based on the performance of only 25 to 35 students, there was substantial class-to-class and cohort-to-cohort variation within schools that might lead to anomalous rankings. As one principal said, “You know, every now and then you get that class.” Similarly, Ms. Dale described a statistics course in which she learned that:
Tests like the CST are designed for districts to evaluate themselves and big schools to evaluate themselves. It's not even really designed for little schools. And so, when you start bringing it down to a teacher, you have to be very, very careful because your sample isn't big enough.

Even though these principals felt that overall there should be a strong correlation between teachers’ TLF ratings and teachers’ AGT rankings, they believed the AGT ranking needed to be used with caution at the teacher level.

Even with that cautionary approach, these principals did not dismiss the AGT as a measure. In fact, they were often adamant that some teachers were effective and some were not, and that this difference was reflected in test scores. As Ms. Dale said:

Some teachers bring test scores up and some don't. . . . Every single year some teachers will blame the kids, but if those kids were in a different room . . . you'd see a different pattern.

These principals, like Ms. Dale, expected there would be cases where the measures might provide different signals about teacher quality. They believed that outcomes matter, and were equally committed to looking at all available evidence—AGT rankings, TLF ratings, tests, classwork, homework—to figure out how to improve learning. However, they were not ready to relinquish their own professional judgment in making decisions about teacher quality.

Like several other principals described in Chapter 6, Ms. Dale systematically placed challenging students who were difficult to reach with teachers she believed were the best teachers, even though, as a result, she expected these teachers to have AGT rankings that might not fully speak to the quality of their instruction. It is also possible
that what Ms. Dale perceived but did not fully articulate is that when some students were initially so low scoring, even if they made gains, these gains might not be reflected on the CST, which was still too hard. Ms. Dale appeared to be concerned that in situations like this, privileging AGT rankings could underestimate and discourage teachers, including teachers who in her professional judgment were good teachers.

Ms. Kim suggested the AGT and the TLF seemed to be correlated about “80 percent of the time,” but that there were cases where the AGT rankings were surprising. Ms. Kim stated that in her experience, these were usually cases where people she perceived as weak teachers had surprisingly high AGT rankings. By way of example, she said:

We had this one teacher, he was terrible. He was terrible. Honestly, I think [he had a high AGT ranking] just because the kids read all the time, and he had smart kids. They kind of taught themselves. Or they were able to really focus on the test. I don't know, because he was such a terrible, terrible teacher.

In this case, Ms. Kim didn’t want this teacher to get the impression he did not need to improve because he had a high AGT ranking. Like Ms. Dale, Ms. Kim referred to her professional judgment, and noted that, while results matter, good teaching goes beyond what appeared to be captured by the AGT rankings.

Ms. Island, the principal in the lowest-performing school in the sample, was surprised by many of the AGT rankings of her teachers. In particular, she stated that many of her teachers had very low rankings, but the ones who had high rankings were always the teachers who had the gifted cluster in her school. This made her suspicious of the rankings. She said: “But they’re not teaching. The kids already have the skills.” Just
as Mr. Isaac felt students at the top of the score distribution would have trouble demonstrating growth on the tests used for accountability purposes, Ms. Island questioned whether classes composed primarily of students scoring far below grade level would also have difficulty demonstrating growth on test-based measures, regardless of how well teachers taught.

Consistent with this suspicion, Ms. Kim also noted that, in cases where students might be nonrandomly assigned to teachers, it might be unfair to hold these teachers accountable based on AGT rankings, and particularly so when the children in the class were students who were initially very low scoring or challenging to teach. She explained:

A principal or somebody can say, “It's about improvement, it's about improvement.” But if you have a classroom full of students who are scoring as “Far Below Basic” and you don't have role models for them, it's hard to move those kids. Really, really hard. Even if you're great, it's hard to move those kids. Her comments suggest skepticism that the AGT can systematically and adequately and accurately account for the challenges posed by educating some groups of children, particularly given current patterns of segregation across school sites. Though Ms. Kim did not discuss it, there is a possibility that the tests were just too hard for some students, such that even when they learned quite a bit, it would be difficult for them to show much growth on a fixed-form test that is too difficult. This could explain why she found it extremely difficult to “move the scores” of students who are “Far Below Basic.”

These accounts of principals in medium- to low-scoring schools provide a stark contrast to the comments of principals in some of the highest-performing schools in the sample. As noted in Chapter 4, principals in some of the highest-scoring schools
suspected that a “ceiling effect” might limit the ability of their teachers to get above-average AGT rankings. Mr. Isaac described teachers who had expert knowledge, served as instructional coaches, and provided professional development, and appeared by all observational measures to be outstanding teachers, yet nevertheless consistently had average or below-average AGT rankings. For Mr. Isaac, this posed a conundrum: when measures suggest different realities about teacher quality, which should he and his parents and students trust? As he observed:

This goes back to this whole issue of validity again. Here you’ve got this teacher who, on the one hand gets this wonderful feedback from you acknowledging all these things she’s done relative to the framework and then she gets these lousy scores. What has authority—the AGT ranking or your rating—in terms of what good practice is?

Mr. Isaac reflected that cases like this prompt questions about what good practice is and how we can tell when it occurs: do observational ratings need to be revisited to be brought in line with the test-based measures, or might there be a limitation to the test-based measure?

**Effect of Discrepancies on Principals’ Judgments**

When faced with discrepant cases, many principals reported their response would be to go into the teacher’s classroom more frequently to get confirmatory evidence related to the rating they had assigned the teacher using the TLF. As principals with this response explained their logic, several patterns played out. Some were inclined to assume that perhaps there had been an error in their rating, and looked for further evidence
consistent with what they assumed was indicated by the AGT ranking. Other principals tended to trust their TLF ratings and were more inclined to discount the AGT ranking, and look for factors that would explain why the AGT ranking might not fully capture the story of teacher quality.

Why would a principal question his or her TLF rating when faced with AGT rankings that appeared to contradict their judgment? Principals cited several factors, including weaknesses in the observation process and a tendency to be too harsh as a rater. For example, Ms. Rosa said that, faced with discrepant cases, her tendency would be to visit classrooms more often, to see if there was evidence she had missed or some other critical aspect of instruction that was driving results but that was not captured by the TLF rubric. She mentioned the training and calibration process she had gone through when learning to use the rubric, and said that the calibration process heightened her awareness that there might be areas or aspects of practice for which she was not as accurate as she might need to be because she might jump to inferences for which she did not actually collect evidence, or perhaps more likely, she might not have noticed critical evidence that might challenge her to revisit her rating.

One principal noticed that the structure of the AGT formula inadvertently rewards teachers of students with more human and financial capital at home. For example, two students could score the same in fifth grade. But suppose the first student had parents who helped with homework and emphasized the value of education. The second received no help or support out of school. Even if these students were demographically similar, the first student was likely to make more academic progress before the sixth-grade test, and thus parental input will show up in the AGT ranking of the teachers of that first child. In
contrast, the teachers of the second child may look relatively less effective, because that child’s learning is not enhanced or reinforced by parental support and expectation. To the extent that parents with more initiative or higher levels of expectation may pursue certain schools or certain teachers within schools, it is likely that students with support were not randomly distributed across schools and classrooms, and principals suggested this could bias AGT rankings downward for teachers who taught students living in the most extreme adversity.

While in theory the AGT rankings control for differences associated with parent income, Ms. Kim was emphatic that not all families that live in poverty are the same. In her community, families with limited means were disproportionately immigrant families with high levels of education and aspiration, compared to communities that looked demographically similar on paper, but which in reality were composed of families that had lived in poverty for several generations and had much less human capital to support their children’s learning.

In an earlier chapter, Mr. Isaac suggested that teachers in his comparatively high-scoring magnet school might have an advantage with AGT rankings over some neighboring schools, despite having similar demographics, because even though his students were characterized by the same poverty metrics, their parents had more social capital and capacity to support the learning of their children than did the parents of some children in other higher-poverty schools. If true, Mr. Isaac’s hypothesis would suggest that the AGT rankings might be negatively affected in some schools that serve students living in the greatest adversity or in multigenerational poverty. Given that magnet schools, residential segregation, and some element of school choice may result in
nonrandom sorting of families and children across schools, this is a plausible concern. In contrast, the TLF ratings give principals more nuanced information on teachers and classrooms that are necessary to put the AGT rankings in context.

When faced with tension created by the high stakes associated with tests on the one hand, and their professional judgment on the other, many principals defended practices in their schools on instructional grounds, in spite of test-based evidence. For example, in his school, Mr. Roland worked hard over a period of years to strengthen the capacity of his grade-level teams to work collaboratively on professional learning and improvement of teaching. These teams reviewed student performance data and worked to develop what they felt were pedagogically and developmentally appropriate pacing guides for their students. Just as this work was completed, the district increased its emphasis on the periodic benchmark assessments, which were supposed to be predictive of performance on the end-of-year summative assessments that were used to compute the AGT rankings.

The benchmarks assessments were meant to drive improvements in scores and create some urgency around pacing and mastery. Mr. Roland’s challenge was that these periodic assessments became a de facto guide as to what teachers would teach and when they would teach it. The benchmark tests specified interim points at which specific standards would be assessed. In order to do well on these benchmark assessments, teachers had to follow that schedule. Mr. Roland noted that when one of his grades did not do well on the periodic assessment, it was because the team had not yet taught the related standards. However, failing to do well triggered a call from the “central office,” which he said felt like pressure to realign instruction around the periodic assessments,
rather than the thoughtful yearlong instructional plans he said his teams had developed. He was committed to supporting his teachers’ professional judgment, and concerned that reorganizing around the benchmark schedule would incentivize bad practice. Because his school’s results had improved quite a bit during his tenure, he was comfortable pushing back. Had his school not demonstrated substantial gains, he might have had a very different conversation.

In sum, while the AGT rankings were generally perceived by principals to be consistent with what they expected based on their TLF ratings, in some cases, they were not. When they were not, principals generally responded by reviewing what they knew about the AGT and TLF measures and reflecting on or repeating their standards-based observations, to review whether there was some indicator of quality they might have missed in their initial evaluation. They also reviewed other available data, including interim assessments, classroom assignments, and class composition. None of them automatically deferred to the AGT ranking when it contradicted the inferences they had made based on observations and professional judgment. Instead, they sought to contextualize the AGT rankings by understanding them in reference to a broader review of available measures and data.
Principals sit at the intersection of process and outcome measures of teacher quality. They must draw on their professional knowledge and understanding of learning to foster growth and improvement in teachers’ practice, and they must reconcile the practice they see on a daily basis with the story told by external, test-based measures of student learning. They need to reconcile rankings based on an aggregation of student scores with the particulars of the interactions they observe between students and their teachers. They are challenged to follow a logic of accountability and evaluation, even as their close examination of the particulars of individual teachers and classrooms clouds the clarity implied by the AGT rankings. Because of their proximity to the work of teaching, principals are uniquely positioned to offer insight on factors that both attenuate and strengthen the correlation between these measures. They can also inform where these measures fall short and where they add value to the essential and all-important work of improving instructional practice, and by extension, student learning. And just as importantly, their deep knowledge of local context and teaching practice offer them unique insight about where and when policies are likely or unlikely to achieve their intended goals and why.

In order to understand the implementation challenges that shape how policies are operationalized on the ground, we have to go beyond cross-sectional data. Robert Yin (2009) argued that case studies are uniquely valuable for illuminating the complexity of how and why people operate in their own ecosystems—in other words, how policy directives are mediated by those tasked with implementing them. Or, as Mintzberg (1973)
might say if he wrote about contemporary principals, we learn how and why principals use tools like the AGT rankings and the TLF framework in shaping the day-to-day practice and culture of the schools by examining them individually, not by scanning “flocks” of them in district data. Without this close examination, we can’t adequately anticipate whether and why policies might succeed, or understand why they might fail.

As discussed in the introduction, copious cross-sectional data suggest that most students are still performing below aspirations and the quality of schools is not improving on average at the pace policy makers and the public would like. At the same time, evidence from schools suggests that most teachers are receiving satisfactory evaluations, and our evaluation tools and processes seem to be inadequate in identifying teachers who are able to generate substantial improvements in learning from those who do not. Policymakers, the general public, and local district leaders have a strong policy interest in strengthening teaching practice in ways that improve outcomes, and in ensuring that principals accurately evaluate teachers and provide them with the supervision and feedback they need to ensure every teacher is improving. The paradox of policymaking, of course, is that, as Milbrey McLaughlin (1991) wrote, “policy cannot mandate what matters” (p. 187). Policy also cannot mandate what happens. Policymakers may identify the problem, but administrators in schools, with teachers, are the ones who have to solve it and the ones with the complex knowledge of their systems to know what will and won’t work in their particular context.

Policymakers, whether federal or local, also cannot just order principals to provide better evaluation and support, precisely because principals may not have the means, resources, capabilities or leverage to do so. As Cohen and others (2007)
explained:

The policy makers who define problems and devise remedies are rarely the ultimate problem solvers. They depend on the very people and organizations that have or are the problem to solve it. At the same time, those that have or are the problem depend on policy makers or others for some of the resources—ideas, incentives, money, and more—that may enable a solution. (p. 522)

In other words, policymakers want to create pressure to improve outcomes by improving teaching practice. They think the key to doing so is to provide stronger evaluation and supervision. For the policy to succeed, however, principals need to have the capability and the intent to conduct more effective evaluations and provide high-quality supervision and feedback, and the tools they are given to use must work as intended. Without that capability, they cannot achieve the policy goals, and may actually subvert them in an effort to avoid sanctions, if and when policy makers attempt to hold them accountable.

We are in the midst of a national policy era in which the focus is on accountability, and in which the prevailing logic says that if we develop empirical measures based on cross-sectional data, and then hold our educators responsible for improvement, student learning will improve. For the most part, contemporary national education policy initiatives are still focused on test-based accountability, and thus give short shrift to management and evaluation of knowledge for teaching and learning. As Jal Mehta (2013) wrote, the logic or technology of standards, assessments and accountability bring with it certain predictable virtues: it allows the state to track the performance of students and schools; it exerts pressure and focus attention towards collective goals; it can bring attention to the
ways in which schools are not achieving what we hope. But it is a limited technology: it has no mechanism to increase the skill of teachers and schools; consequently, it has made very limited progress in closing achievement gaps or improving America’s educational standing in the world. (p. 267)

The theory of change behind this approach is based on urgency and pressure: tell educators to perform or be sanctioned, and they will feel urgency and work harder. This logic was fully espoused by only one principal in the sample, a very young principal in a low performing school, where she saw her job as primarily one of creating urgency for improved outcomes.

To some extent, the LAUSD’s Educator Growth and Development Program pilot attempted to address the needs of both policymakers and educators. The pilot paired use of a test-based growth measure with a substantial investment in professional development and support for using the standards-based TLF. If what leaders need to do is “increase the skill of teachers” (Mehta, 2013, p. 267), I would argue that principals in this sample felt the TLF did more to advance that goal than the AGT. Thus, the pilot both tied teacher evaluations to student outcomes and provided teachers with substantive, high-quality feedback about performance they could use to improve their teaching practice.

This pilot thus introduced both tension and support into the work of principals. The AGT ranking introduced tension, because all teachers were ranked individually relative to peers, and some were labeled with red and orange colors and ranked as below average. The rankings focused teachers’ attention on the importance of measuring outcomes as measured by tests, but some principals and teachers felt this was punitive, and all the more so when factors beyond the control of teachers potentially shaped the
rankings, and because the AGT rankings provided no insight into what was “wrong” with the teaching that led to a low ranking. In particular, concentration within schools and classrooms of extreme poverty was a subtext of many conversations. The interviews raised questions about whether, even with good tools and good intentions, educators in schools with the highest levels of family instability and proportions of students in extreme adversity can effectively “teach” their way to better outcomes for all students under the current circumstances.

Ratings based on the TLF also introduced tension, particularly as principals began giving lower ratings to experienced teachers who had previously been rated quite favorably. However, where the TLF differed from the AGT was that the TLF and the associated training and materials provided principals with the support and evidence they needed to feel more confident in their ratings and more effective at helping teachers respond to evidence about the alignment of their practices with best practices in ways that educated them on how to improve. In fact, most principals could not talk about the TLF without talking about improvement.

Policy works when those who implement it embrace the purpose and have the capability to execute it well. To the extent that principals understood and embraced the purpose of the AGT and the TLF, and could use these tools well to accurately and consistently diagnose teacher quality and support continuous improvement in teaching practice, the EGDC policy would likely lead to better student outcomes. In practice, the test-based measure (the AGT) was easy to access and use, but less trusted and valued. In contrast, the standards-based observation process (the TLF) was challenging to implement and use, but highly valued. In some places, principals wondered if the
individual “shaming” associated with the AGT rankings might erode the professional
colleagiality they felt was so essential to improving teaching practice. With one exception,
principals were ambivalent about the AGT ranking as a teacher-level measure, and firm
that, if used, it needed to be used by principals with good judgment and contextual
knowledge. However, they valued the TLF as a tool that reinforced and strengthened
their role as instructional leaders, and enhanced their ability to support improvements in
the practice of their teachers. They felt this way, even as they noted how time consuming
and demanding it was to complete the process. They expected that, if teachers enacted the
effective practices described in the framework, student learning would improve. However,
they also identified several challenges that could call into question their ability to sustain
effective use of the TLF over time. The most salient threats include pressures of
accountability, and the capability over time to sustain fidelity of implementation,
especially in the lowest-scoring, most poorly resourced schools, where the demands on
school-based leaders are extraordinary.

Prominent in the contextual background for all principals in the sample was the
specter of accountability; both principals and teachers were aware that failure to
demonstrate strong outcomes likely could have consequences, particularly in the lowest-
scoring schools, like those led by Ms. Mitchell, Mr. Grace, and Ms. Island. Principals in
high-scoring or improving schools, like Mr. Isaac, Ms. Alana, and Mr. Roland, appeared
to have more agency and more capacity to discount or mediate the potential perverse
consequences of accountability discussed in this study. In a few cases, they were able to
deconstruct the AGT ranking to explain with great sophistication why it might be a
flawed measure of teacher efficacy. In contrast, lower-scoring schools labored under a
label of failure, and with one exception (Mr. Grace), expressed great pressure to improve outcomes regardless of any concerns about measures. However, in the absence of time, knowledge, resources, support, and community capital, teachers in these lower-scoring contexts may not be able to improve their own teaching practice rapidly enough to meet the policy goals of higher scores through more effective practices before being subject to the consequences of accountability.

Indeed, many principals alluded to this when they suggested that even though they strongly believed that teachers need to take responsibility for educating all children and improving outcomes, not all children are the same, and not all children are easy to educate. Cohen and others (2007) suggested:

In such cases, teachers can be expected to reduce risk rather than improve performance because they lack the knowledge required to improve performance, and instruments do not provide it. Under these circumstances, if policies attach consequences to performance, they increase the chance that practitioners will try to reduce risk by complying without improving practice. (p. 536)

For example, teachers may feel under tremendous pressure to improve their AGT ranking in order to avoid sanctions or public embarrassment when the rankings are published in the newspaper. If at the same time it is beyond their capability to improve those outcomes through the desired means (better instructional practice), they may resort to other practices or behaviors that improve scores in the short run, but which are not associated with the kinds of generalized improvements in learning that are the ultimate goal of policymakers. This would represent a perverse effect of AGT rankings.

In fact, many principals provided examples of how this perverse effect can play
out. Some described modifying teaching to focus on the specifics of how content is sampled on the test, so that test items would be familiar. Some described focusing instruction on the standards that are actually sampled on the test, at the expense of other equally important standards that are not sampled. Others described giving up the learning that is engaging and pushes students to deeper levels of understanding, in order to focus on what counts on the test. Some alluded to the possibility of cheating. Some described teachers who try to push “hard-to-teach” students out of their classes. If teachers feel their jobs or reputations are at risk, it is reasonable to expect them to respond in ways that reduce that risk of job loss or damage to their professional reputations, even if these responses are not compatible with the larger policy goals. If the perceived threat of accountability outweighs the capability of educators to deliver—whether because of resource limits or their own lack of skill—policymakers, principals, and teachers get what they need (higher scores), but they may not get what they want (better learning).

This supposition is consistent with other research that suggests scores on many state tests in states with strong accountability systems have gone up substantially in recent years, even as performance on the low-stakes National Assessment of Education Progress has tapered off or remained flat. If learning has truly improved, we would expect gains on NAEP to be commensurate with gains on state tests. It goes without saying that, if accountability pressures in low-scoring schools are more intense, these schools are more vulnerable to this kind of erosion of practice. This could potentially erode the goal of equity, if students in low-scoring schools are disproportionately exposed to practices that raise scores on tests used for accountability purposes, but perhaps at the expense of other valuable learning critical to longer-term success in more
advanced coursework or life beyond school. In fact, in these contexts, rising scores could actually mask continued inequities in the short term, making them harder to address.

With respect to measuring teacher quality, most but not all principals paid attention to both measures, but preferred the Teaching and Learning Framework. To some extent, this was a function of professional knowledge of and value placed on the AGT ranking. While the AGT ranking spoke to the importance of outcomes, the opacity of the formula used to generate the rankings, coupled with the occasional “surprising” ranking, raised questions for many principals about what the AGT ranking actually measured. In addition, principals’ discussion of the AGT ranking suggests a strong need for additional professional development related to the test-based measures, what they actually measure, and what inferences they can support. For example, a striking number of principals referred to the AGT ranking as an absolute measure of growth, not a relative ranking. Those who discussed it as a relative ranking expressed a desire for an absolute measure of student progress, so that any discussion of growth could be benchmarked to some standard.

This suggests that, conceptually, the principals valued the idea of a growth measure that could accurately assess the ability of teachers to generate greater learning. However, as Mr. Isaac put it, many just did not think the AGT rankings were “ready for game time” as consequential measures. Principals preferred growth models to absolute measures of student performance, but persistent skepticism about the tool, coupled with a lack of clarity and limitations on how to use it to improve performance, tempered their enthusiasm and use of the tool in many contexts. And, several of the principals referred to the AGT rankings, like Ms. Dale did, as “old news.” Once principals received the
rankings, they were already historical data and teachers were already well into the process of new plans and working with different groups of students.

Principals were aware that the TLF ratings might also be subject to challenges related to bias and reliability. However, they repeatedly stressed that the calibration, the careful emphasis on matching evidence to descriptors of practice reduced this risk. Many explained that in their minds, the strength of using the TLF lay in part in its transparency. Moreover, the explicit descriptions of practice and matching of evidence had shifted their conversations in powerful ways. Conversations were less focused on whether their summative judgments and feedback were valid, and more on the specifics of practice captured in observations, and on what teachers might do next to improve their practice. The tool enabled them to move beyond judgment to a focus on instructional improvement.

A recent study by Grissom, Loeb, and Masters (2013) found that, while many uses of principals’ time do not appear to predict improvements in achievement, the time that they spend coaching and evaluating teachers is associated with greater test score gains. The principals in the sample appeared to share this belief as well. With respect to supporting growth and improving practice, the principals in the sample overwhelmingly preferred the TLF as a tool for generating improvements in instruction. The specific descriptions of practice, coupled with the use of scripted evidence, were universally described as effective for helping teachers understand their strengths and areas needing improvement, as well as what they could do to lift their practice to the next level.

There were some striking differences across high- and low-scoring schools in how principals used and perceived the authority of test-based measures. In her high-poverty
school, Ms. Mitchell used the rankings and frequent periodic assessments to track progress on test-based measures, and she adjusted feedback using the framework to be consistent with improvement (or lack of improvement) in scores. Observational feedback worked tightly with periodic assessment data in a tight, performance-monitoring and feedback loop. Some other principals in low-scoring contexts described using the periodic assessments to pace learning and track progress toward end-of-year tests and in some cases, they were pleased when periodic assessments suggested more students would do well on end-of-year tests. Most seemed resigned that their scores might not be strong, but they still focused on striving for incremental improvements.

In schools where principals closely tracked test results, there was a tendency to assume strong alignment of AGT rankings and TLF ratings, on the logic that if teaching were better, scores would be better, and that if scores were not good, the teaching probably was not good either. The danger in this assumption, of course, is that if principals expect their TLF ratings to align with the AGT rankings, that might bias their TLF ratings, which in turn could erode their usefulness for improving instruction. Given multiple factors identified by principals that might bias or erode the quality of the AGT ratings, this risk should be considered and evaluated by principals and their supervisors in all contexts. Additional research is needed to explore this phenomenon.

In contrast, some principals, and particularly those in higher-performing schools, took a skeptical stance with respect to the AGT, which shaped how they used and discussed the AGT with teachers. For example, some principals expressed open doubt at the high AGT rankings of certain teachers, whom their own observations suggested were weak. Others pointed out, for example, that the science tests tended to emphasize content
knowledge in a way that rewarded “drill and kill” teaching, while the frameworks emphasized scientific inquiry. One pointed out that the AGT rankings for science teachers used prior-year scores in math and English language arts to evaluate growth in science, and questioned how growth measures for science teachers could credibly be based in part on scores in other subjects. A middle school principal suggested that beginning-of-the-year assessments suggested the elementary CST scores in some of her sending elementary schools were inflated, and wondered if this would bias the AGT rankings of her teachers downward. This questioning created space for conversations about effects on outcomes that fell outside the narrow parameters of test-based accountability.

The interviews indicated another feature of differences across schools. Being a higher-scoring school usually correlated with socioeconomic and racial differences in the student population. Higher scores also gave these schools the privilege of adopting a broader set of learning goals and greater freedom from narrow performance pressure associated with the AGT rankings. Principals in high-scoring schools functioned as “protectors,” mitigating what they perceived as corrosive effects of policies on their schools and teachers, while promoting any positive effects in their school contexts. These principals, particularly in higher-scoring schools, seemed to have the credibility and public support to discount the AGT ranking and make decisions based on professional judgment. For example, several principals in higher-scoring schools observed, to borrow the words of one, that “no one really cares about the AGT.” Others noted that parents sought to have their children placed based on a teacher’s reputation, sometimes in contradiction of “evidence” provided by the AGT. Another noted that the CSTs captured
basic material, and not the “more interesting” work that her community expected. She stated that in her school, parents looked at their children’s work as evidence of performance, and gave less importance to the AGT rankings. She noted that her school could do this because its students were high performing, but lower-scoring schools might feel pressured to emphasize the basics at the expense of anything “interesting.”

In contrast, principals in low-scoring schools, which typically also served more students living in poverty and more students of color, felt more pressure to focus improvement efforts on activities that yielded higher scores. Principals and teachers worked together to identify content and formats emphasized on the tests. Not only were activities driven by performance on the summative tests, but also by the periodic benchmark tests, which, as some put it, also became “high stakes,” because if your students did not do well on interim benchmark tests, you were “likely to get a call from central office.” The effect was to align tightly the scope and sequence of curriculum with the periodic benchmark tests and the standards addressed on each of these interim tests, regardless of the students’ level of mastery or progress.

Though principals felt the TLF was very powerful for stimulating professional reflection and motivating improvements in instruction, without exception they expressed the concern that it would be difficult to implement with fidelity, at scale. The prescribed observation cycle was necessarily time consuming, between pre-conferencing, scripting, tagging of data, and planning for improvements. To use the tool well and have reliable ratings, principals felt they needed to conduct several observations. Having fewer instances of observation, in their minds, reduced the reliability of ratings because they might not be able to gather sufficient evidence related to all indicators in the framework.
within a single instance of observation. While all lauded the value of the process, they all also questioned how they would be able to use the framework well, on top of all their other duties. In describing their experience, several principals used the word “overwhelmed.”

**Unique Challenges Associated with Concentrated Adversity**

Two of the most important takeaways from this study were actually two factors that did not fit neatly into the findings. First, it was impossible to read the interview transcripts without being struck by the overwhelming complexity of the work of these principals, and the incredible insight, sophistication, and investment of time it took to bring coherence and focus to a policy context characterized by multiple initiatives and competing incentives. For the most part, these principals sought to focus their work by identifying results or measures that they felt were central to the work of their schools. They then worked opportunistically with those tools, and in particular, the TLF, to support conversations and motivate work that would improve student learning.

Second, the interviews took place in a diverse sample of schools, from three that educated extraordinarily successful students, to three that supported a desperately economically stressed, racially segregated, and low-performing population. In the high-performing schools, models could be applied with some integrity, and when leaders discounted or distrusted certain tools or their uses, the tools could be disregarded without ill consequences. In stark contrast, the work of the three schools that struggled most was shaped by basic challenges related to safety and student turnover. One school with a highly mobile population had 50% turnover in enrolled students during the preceding
academic year, rendering laughable the notion of consistently and carefully following a pacing guide. As one principal commented, students need to learn fractions before they can learn algebra, regardless of what the pacing guide says.

In several of the lowest-performing schools, the interview was repeatedly interrupted by, for example, concerns related to children who had run away, children who had been mugged on the way home, children who were in difficult custodial disputes or suddenly became homeless, or children who had come to school after getting into a fight. In these contexts, the school was evidently a source of stability and safety that out of necessity took priority even over academic learning. In these contexts, principals spoke about trying to keep teaching and learning front and center, but the reality of daily demands made it challenging to do so. Students needed much more intensive support and responsive teaching than in some of the higher-performing schools, and the demands on teachers were much more intense. Because some of these schools served a high proportion of students who were performing far below grade level and whose lives were so turbulent, it was easy to wonder if the full contribution of teachers to these children could be fully captured on the grade-level tests used for accountability purposes. If they were, would these gains adequately speak to the enormity of the challenge the students of these teachers faced?

In addition, because these schools served highly disadvantaged students, they also had lower test scores. Significantly, state and federal determinations of performance are framed in normative language, and judgments of low quality are reinforced in the red labeling that emphatically identifies them as low performing and as “focus” schools for intervention. This negative narrative hung over these schools. The implication was that
these schools were failing children, and needed strong accountability for outcomes and a stronger sense of urgency in order to turn around. It was the job of schools and teachers to break the cycle of poverty, and if they did not, they “failed.” In some of these schools, principals used the AGT rankings forcefully to push teachers to fix the problem of poor instruction. Yet, as Cohen and others (2007) noted, if educators are to improve weak schools, it also makes sense for policy to offer the assistance, encouragement, and trust that they need to do the job (p. 541). These schools seemed trapped in a paradox: on the one hand, strong policies pressured principals and teachers to improve performance. On the other, the district was not making the necessary resource investment nor was the City of Los Angeles providing the supports for children and families that might actually enable these high-poverty schools to achieve the desired outcomes for students. Accountability pressure, combined with insufficient support in the face of grinding poverty, demoralized staff and eroded the integrity of their efforts to improve performance.

When educators feel under pressure and are held accountable to improve performance, but lack the professional expertise or opportunity or capital they need to do so in ways that are called for by policy, they may behave in ways that meet compliance requirements but do not actually lead to longer-term, sustainable solutions that might counteract growing inequity. One additional finding of this study is that if standards-based observations are to have the same powerful effect on improving instruction in high-poverty schools that they do in lower-poverty schools, these high-poverty schools may need much more support and more staffing than lower-poverty schools in order to
support successful implementation of the TLF and reap the expected benefits (improved student learning). This proposition should be explored in subsequent research.

Finally, recent research (Jackson & Bruegmann, 2009; Leana, 2011; Sun, Loeb, & Grissom, 2015) suggests that teacher efficacy is not so much an individual quality, but a collective quality. A lower-capability teacher is improved if he or she is in a school with many effective teachers; inversely, in contexts that are corrosive, unstable, or struggling, even a high-capability teacher will struggle more than she or he would otherwise. Several principals seemed acutely aware of this dynamic. Mr. Roland organized his math teams so that, in effect, their AGT rankings were group measures, not individual measures. He did so because he realized that different teachers were more effective at teaching certain concepts than others, and by teaming and sharing students, he was able to leverage individual teachers’ strengths for the benefit of the group. Mr. Isaac discussed the need to foster accountability of teachers to each other for improving performance, which school-level AGT rankings do. In two of the highest-performing schools in the sample, teachers’ work was structured so that they had opportunities to give and receive feedback with colleagues, using the TLF. In contrast, the principal in the lowest-scoring school, Ms. Island, had introduced looping, requiring teachers to take students all the way from kindergarten through third grade, specifically to heighten individual accountability. The interviews with principals hinted at a collective nature to school improvement that is not yet adequately contemplated in supervision and evaluation models, and perhaps not uniform across school contexts. While principals in higher-scoring environments can mediate the effects of evaluation tools in ways that facilitate collaboration, accountability pressures in some contexts put the focus on individual demonstration of progress by
individual teachers. Future work should continue to explore the relationship of a culture of professional collaboration to both student outcomes and successful implementation of evaluation and supervision models.

Conclusion

This study, like any qualitative, small-scale study, has limitations that prevent generalization from the particular across all contexts. However, close analysis of the experiences and views of a sample of principals highlights aspects of policy and implementation that can be tested or explored in subsequent work, and which can inform implementation of policies and practices that seek to improve instructional practice.

First, this study suggests that principals’ ability to implement policy is a function of their knowledge of the tools they are given to achieve policy goals, the value they place on these tools and goals, and their possession of the capabilities, time, and resources to implement the policies in the context of the daily stresses and demands of their school environment. Policy will flourish or founder in different places, depending on the mix of these factors in different contexts. If, for example, principals do not feel an evaluation tool accurately and consistently captures teacher quality, they will be indifferent or inconsistent in their use of that tool.

It is also clear that high-scoring schools and low-scoring schools appear to play by different rules, because higher-scoring, higher-capital environments can mediate the effects of policies in ways that schools that are lower-scoring and which typically serve disproportionate numbers of students living in poverty cannot do. Leaders in highly stressed environments just do not have the powerful agency that leaders in less stressed
environments have. To achieve instructional goals, these stressed schools may need far more support for students’ needs and for administrative responsibilities than currently is available. For example, successful implementation of the Teaching and Learning Framework may take much more time to implement at scale than is currently acknowledged, and this is particularly true for schools that serve very disadvantaged populations. Schools that serve disproportionate numbers of students in poverty may need higher staff-to-student and administrator-to-teacher ratios to effectively implement the TLF and support improvements in learning.

Weak schools tend to stagnate, not because they do not desire to improve, but because they lack the capability to do so on their own. Once schools have been labeled failing, punitive accountability models often further disable them by making it increasingly difficult to attract and retain the teachers with the capability and experience necessary to make these schools strong. The negative narrative of failure around these schools also discourages parents with options from sending their children to these schools, which further exacerbates the social isolation and economic deprivation against which these schools struggle. The extreme adversity in which many of their students live makes it difficult to maintain instructional focus and momentum, and less likely that parents have the resources, connections, and time to support their children. In these schools, many principals seemed to seize on the TLF, in part because it allowed them to pivot from a focus on ratings, rankings and low test scores to a focus on continuous improvement of teaching practice.

This study suggests that the Teaching and Learning Framework has powerful potential as a tool for instructional improvement, if well supported and well implemented.
The clarity and specificity of the tool, coupled with an emphasis on the use of evidence to support inferences, enabled principals to conduct powerful conversations about teacher practice, and in particular, the relationship between what teachers say they do (espoused practice) and what teachers actually do when they are in classrooms with students (enacted practice). Across the board, principals stated that the TLF enabled them to give specific, actionable feedback in ways that had proved difficult under the previous evaluation model. In sum, while the impact of the AGT rankings on teaching practice is unknown, the level of principals’ commitment to the TLF process suggests that in those schools that are able to implement the TLF with fidelity, it may shape practice and improve outcomes.

Further research might explore how to enable principals and teachers in low-scoring schools to have the time and other resources necessary to use the TLF effectively to support professional learning and improved student outcomes. Supervising teachers effectively takes substantial time, emotional strength, and persistence. These factors are precious commodities in highly stressed schools. Without strong support from the district and greater instructional knowledge and skill, it may not be realistic to expect leaders, particularly in those high-poverty contexts, to sustain the necessary energy to implement this work over time.

Interestingly, principals described the TLF as a tool for both building teacher capabilities and documenting the poor performance of teachers who need to move to other professions. These dual purposes confirm that principals play a pivotal role in ensuring every child is taught by a skilled teacher. Mr. Isaac’s comments in Chapter 7 suggest that schools get better not because they have better measures to tell them how
they are doing, but because districts and school building leaders develop the capacity of
their professionals to evaluate evidence and act on it collaboratively in ways that lead to
improved learning. Mr. Isaac explained that his role as a principal was shifting from a
focus on rating and prescribing, to a focus on mission and results, while building the
capacity of his staff to achieve results. His work, like the work of other principals,
increasingly involved questioning and challenging and providing support, as teachers
developed their own capacity to support better outcomes for students. This reinforces the
importance, moving forward, of continued research to inform strategies for developing
and supporting principals as they engage in formative work with their teachers.

Finally, while current reform efforts position the principal as the leader of
instruction, most principals in this sample suggested a slightly subtler role for themselves.
First of all, principals in the sample didn’t just use the TLF in private conversations to
give individuals feedback. Instead, they used the TLF in collaborative and public ways to
build a shared understanding of practice and to foster collegial conversations around what
good teaching looks like in the classroom. Second, they didn’t determine or define
teacher quality purely in terms of the AGT ranking or other “objective” measures of
teacher quality. Instead, they emphasized the critical role of professional judgment in
reviewing a variety of data and diagnosing teacher strengths and weaknesses. And finally,
these principals positioned themselves as mediators or protectors of their schools,
students, and teachers by mediating the impact of external policies and protecting
improvement efforts from external disruptions. Taken together, these qualities suggest a
theory of school improvement that is animated less by sanctions and accountability than
by careful and persistent, data-rich development of professional capability.

H1: In discrepant cases, principal confidence in the TLF ratings they assign will be mediated by

- Principal level of experience as an administrator.
- Principal understanding of the AGT rankings, what they measure and how they are generated
- Principal understanding of the alignment between what is tested and what is specified in the standards

H2: The value or weight principals place on the AGT measure as an indicator of teacher quality will be mediated by their knowledge of factors that potentially threaten the validity of AGT-based inferences about individual teachers such as:

- Staffing patterns within and across subjects
- Distribution of student performance within the building
- Alignment between how skills and concepts are tested and how they are taught
Appendix B: Method of this Study

<table>
<thead>
<tr>
<th>Study Phase:</th>
<th>Activity:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Designing and Defining</td>
<td>1. Develop propositions</td>
</tr>
<tr>
<td></td>
<td>2. Select cases</td>
</tr>
<tr>
<td></td>
<td>3. Develop protocols</td>
</tr>
<tr>
<td>2. Preparing, Collecting and</td>
<td>4. Conduct first interviews, and write an individual case report after</td>
</tr>
<tr>
<td>Analyzing</td>
<td>each interview</td>
</tr>
<tr>
<td></td>
<td>5. Preliminary review of data to yield preliminary patterns, constructs</td>
</tr>
<tr>
<td></td>
<td>6. Play preliminary data back to interviewees, EGDP leaders</td>
</tr>
<tr>
<td></td>
<td>7. Revise theory and protocol (e.g. go back to Phase 1 before completing</td>
</tr>
<tr>
<td></td>
<td>second round of interviews)</td>
</tr>
<tr>
<td></td>
<td>8. Conduct second round of interviews, and write an individual case report</td>
</tr>
<tr>
<td></td>
<td>after each interview</td>
</tr>
<tr>
<td></td>
<td>10. Modify theory develop policy implications</td>
</tr>
<tr>
<td></td>
<td>11. Write cross-case report and play back to key stakeholders</td>
</tr>
</tbody>
</table>

---

4 Adapted from COSMOS Corporation, Yin, 2009.
Appendix C: Table of General and Specific Questions

What is teacher efficacy as instantiated in the AGT (an outcome measure)?
- What does it measure?
- What it looks like?
- How you talk about it?
- What evidence does it use/provide?
- Do you have any concerns about the AGT?
- How well does it capture what you feel matters? (Potential threats to validity)
- How does it resonates/is it perceived in the building?

How is Teacher Efficacy instantiated in the Teaching Framework (a process measure)?
- What does it measure?
- What does good teaching look like, as captured by the TF?
- How do you talk about it in the context of the framework?
- What evidence do TF ratings rely on?
- Do you have any concerns? (Potential threats to validity)
- How it resonates/is it perceived in the building?

Would you expect there to be a relationship between TF ratings and AGT rankings?
- If so, what would you expect it to be?
- Do they tell the same story?
- Do they measure the same thing?
- How are they related?
- Would you expect one to be predictive of the other?
Appendix D: Interview Protocol (Semi-Structured Interview)

Warm up questions:
1. First, I'd like to ask you a few questions about your role. What do you like about being a principal? What do you feel is most challenging? What is the most important part of the job to you?

Understanding ratings based on value-added scores:
2. I'd like to ask you about when you first heard about value added scores for your teachers. How did you hear that you would be getting this data?
3. When you first heard about the value added scores, what was your understanding of their purpose? [PROBE: Has your understanding of their purpose changed since then? How would you describe their purpose today?]
4. When you first saw the value added scores for your teachers, what was your reaction? [PROBE: Did they match what you thought they would be? Where there any surprises? Could you give me some examples? How did they differ from your ratings of teachers? How do you explain those differences? Ask for examples where principal’s rating was higher and where it was lower than value-added score]
5. What do you think of the value-added scores? In what ways are they useful to you? In what ways are they not?
6. If you looked at AGT scores before conducting the observation, what impact if any did having access to these measures have on how you approached observations?

Scenarios:
7. Present the principal with each. For each scenario, ask: a) How would you interpret this case? b) How do these results make you feel about the relationship between rankings based on student gain scores and ratings based on principal observations? c) Where there were discrepancies, what explains these discrepancies? Why?
   Scenario 1: The principals’ observational ratings for Ms. Jones suggest that she is a strong teacher, in the top 20% of all teachers. The value-added scores for Ms. Jones indicate that she is an exceptional teacher, in the top 5% of all teachers.
   Scenario 2: The principal rates Mr. Smith in the bottom 20% of teachers. The value added scores place him in the bottom 5% of all teachers.
   Scenario 3: The principal rates the Mrs. Peters in the middle 10% of all teachers. The value added scores place her in the top 15% of teachers.
   Scenario 4: The principal rates the Mr. Cook at the in the top 20% of all teachers. The value added scores place Mr. Cook at the in the middle 10% of all teachers.
8. Suppose a principal (Principal “A”) learned that the ratings she gave her teachers were lower than those received by other teachers in the system with
the same value-added scores. How would you explain this? Do you think this would affect how she performed future observations? If so, how and why?

9. Suppose another principal (Principal “B”) learned that the ratings she gave her teachers were higher than those received by other teachers in the system with the same value-added scores. How would you explain this? Do you think this would affect how she performed future observations? If so, how and why?

10. Who would you rather be: Principal A or Principal B? Why?

11. Suppose principal “C” learned that she rated some teachers more harshly than one might expect, given their associated AGT scores, while she rated some other teachers more leniently than we might expect, given their associated AGT scores. How would you explain these ratings? What factors might influence ratings, if any?

12. What is the relationship between value-added measures and your authority as a principal?
   a. How does the value-added measure influence your work as a principal? How does it influence your role in the school? How does it influence your relationships with teachers? How does it influence your work with teachers?

**Personal Experience with value-added scores in relation to ratings based on observations:**

13. Do the ratings you gave teachers based on your observations and the ratings teachers received based on student gain scores tell the same story? Please explain.

14. How do you use value-added scores in your work with teachers?

15. How did comparing the AGT scores to your ratings based on observations influence your understanding of AGT measures? Your own evaluations? How did it influence your perception of yourself as an evaluator? How did it influence your understanding of what constitutes good teaching, if at all?

16. Did seeing the value added scores lead you to reassess any of your ratings? If so, why? How?

17. Does this new evaluation model change your role? If so, how?

**Conclusion:**

18. Is there anything else I haven't asked about that you would like to say?
Appendix E: Initial Coding (Etic Codes) Used for First Round of Coding

<table>
<thead>
<tr>
<th>Teaching Framework</th>
<th>What teaching framework ‘measures’; purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-Me</td>
<td>Nature of effective teaching implicit in TF</td>
</tr>
<tr>
<td>TF-et</td>
<td>Role/nature of evidence collected</td>
</tr>
<tr>
<td>TF-ev</td>
<td>Student learning/engagement</td>
</tr>
<tr>
<td>TF-stud</td>
<td>Time to implement</td>
</tr>
<tr>
<td>TF-time</td>
<td>Concerns about TF</td>
</tr>
<tr>
<td>TF-concerns</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Academic Growth over time</th>
<th>What the AGT measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGT-me</td>
<td>Effective teaching</td>
</tr>
<tr>
<td>AGT-et</td>
<td>Student learning</td>
</tr>
<tr>
<td>AGT-stud</td>
<td>Concerns about AGT</td>
</tr>
<tr>
<td>AGT-concerns</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement Issues:</th>
<th>Reliability (e.g. consistency of measurement’ whether would get the same result if sampled another time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>Accuracy (of measurement, usually with respect to FT rating)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Validity</td>
</tr>
<tr>
<td>Validity</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Correlation of AGT and IF</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Alignment</th>
<th>Alignment of Testing and Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Align-test</td>
<td></td>
</tr>
<tr>
<td>Align-instruc</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effects of measurement</th>
<th>Effect on practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect-teaching</td>
<td></td>
</tr>
<tr>
<td>Effect-culture</td>
<td></td>
</tr>
</tbody>
</table>
### Appendix F: Case Study Tactics for Four Design Tests

<table>
<thead>
<tr>
<th>Type of Validity</th>
<th>How I will address this in my study:</th>
<th>Phase at which I will address this:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Construct validity</strong></td>
<td>✓ Constructs that are the focus on interviews are developed from the literature (AGT, FFT, alignment of test and curriculum, alignment of curriculum and standards, principal understanding of measurement principles)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓ Transcripts of interviews sent to all interviewees, and all interviewees invited to respond or append.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓ Debriefing and discussion of interviews with EGDP administrators after first round of interviews (two meetings with head of EGDP, one with a Deputy Supt., one with the LAUSD Talent Management Team)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓ Sharing of preliminary written report with EGDP and LAUSD Talent Management Team between first and second round of interviews</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓ Sharing of second preliminary report at close of second round of interviews, with follow up discussion with EGDP administrators</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓ I will share a draft of the final paper with principals for comment.</td>
<td></td>
</tr>
<tr>
<td><strong>Internal validity</strong></td>
<td>✓ Pattern matching (Yin, 2009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓ Explanation building</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓ Address rival hypotheses</td>
<td></td>
</tr>
<tr>
<td><strong>External validity</strong></td>
<td>✓ Interviews of 20 principals, to facilitate replication logic (e.g. seeing if empirical data from multiple cases supports analytic generalization (Yin, 2009)</td>
<td></td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td>✓ To ensure the reliability, or specifically, to minimize errors and biases, I will employ an interview protocol, so ensure that each interviewee has an opportunity to address the same questions, and that a subsequent interviewer could repeat the same procedure, and ostensibly gather the same evidence at the same sites.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓ At the end of each day of interviews, I will write a research memo summarizing key themes from the interviews.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓ I will write and append to the dissertation a memo in which I document my own experience with respect to tests scores, school leadership and teacher evaluation,</td>
<td></td>
</tr>
</tbody>
</table>

5 Adapted from Yin, 2009, p. 41
and reflect on how this experience might shape my interactions with interviewees.

I will develop a case study database, which will include transcripts of interviews, case study narrative notes written after each day of interviewing, etic and emic codes used in analysis of data, any tables created during the process of analysis.
Bibliography


187


U.S. Department of Education (2010). Overview information; Race to the top fund; Notice inviting applications for new awards for fiscal year (FY) 2010. *Federal Register, 75*(71).


