Evaluating for Improvement: Perceptions of Tennessee's Teacher Instructional Growth for Effectiveness and Results (TIGER) Observation Model

Erin Imon Gavin & Doug Magee
Vanderbilt University
Ed.D. Capstone
May 2019

ABOUT THE AUTHORS

Erin Imon Gavin serves as the education program director at the McKnight Foundation in Minneapolis, Minnesota. She oversees the implementation of McKnight's strategies to prepare, recruit, and retain diverse, effective educators and to elevate the voice of families and communities in discussions of educational policy and practice. Gavin began her career in the classroom as a teacher and literacy interventionist.

Doug Magee is the Associate Head of Ensworth High School in Nashville, Tennessee. Previously, he served as the Dean of Curriculum Innovation and Design at The Church Farm School just outside of Philadelphia, Pennsylvania. His work focuses on improving school climate and culture through human-centered design with the primary goal of fostering deeper learning for students and teachers.

ACKNOWLEDGEMENTS

We wish to acknowledge the support and guidance of a number of individuals who assisted in the development and refinement of this capstone project. These include Erin O'Hara and Jason Grissom, Ph.D. of the Tennessee Education Research Alliance, Peabody faculty members Claire Smrekar, Ph.D. and Will Doyle, Ph.D., Seth Hunter, Ph.D., of George Mason University, Kevin Schaaf at the Tennessee Department of Education, Courtney Bell, of TN SCORE, and John Campbell, Ph.D., President of the Board, AIMS TIGER Evaluation System of Tennessee.

Although we've promised them anonymity, we extend our gratitude to the countless teachers, school leaders, instructional coaches, and central office administrators in participating study districts who welcomed us into their communities and shared their perspectives on teacher evaluation, feedback, and instructional growth.

Finally, we appreciate the (seemingly infinite!) patience and support of our spouses, families, friends, and colleagues over the course of the Ed.D. program.

Table of Contents

EXECUTIVE SUMMARY

In recent decades, research demonstrating both that teacher quality significantly influences student achievement and that teacher effectiveness varies considerably has stimulated a number of reforms aimed to improve the productivity of America's teachers. Teacher evaluation has emerged as chief among these reforms (Howell, 2015). Spurred in large part by the Obama administration's Race to the Top priorities and Tennessee's First to the Top Act, Tennessee counts itself among the early adopters of teacher evaluation reforms. While the majority of Tennessee districts use the state-developed Tennessee Educator Acceleration Model (TEAM) to gather observational data on teacher effectiveness, a small number of districts use state-approved alternative observation models.

Staff at the Tennessee Department of Education (TDOE) are preparing to review and revise the state's teacher evaluation policies in an effort to strengthen the connection between teacher evaluation and teacher improvement. TDOE leaders have expressed interest in whether features of the state's alternative evaluation models may facilitate greater improvement than the more commonly used TEAM evaluation rubric. In particular, the annually administered Tennessee Educator Survey suggests that teachers in districts using the Teacher Instructional Growth for Effectiveness and Results (TIGER) observation model may have more positive perceptions of evaluation as a tool for improvement.

At the request of the Tennessee Education Research Alliance (TERA), an evaluation and research center at Vanderbilt University which collaborates with TDOE, we undertook an in-depth examination of the TIGER observation model.

Our analysis considered the following questions:

1. Is there a difference in teacher perceptions of evaluation in TEAM districts as compared to districts using an alternative observation model?
2. How do district policies and school practices combine to influence patterns of implementation across districts using the TIGER observation model?
3. How do school and district contexts influence perceptions of teacher evaluation in districts using the TIGER observation model?

These project questions led us to identify trends across evaluation contexts, examine policies, perceptions, and practices which shape teacher evaluation and feedback within the TIGER model, and document variation in implementation within and across a set of TIGER districts. We employed a mixed-methods study that explores the relationship between teacher evaluation models and teachers' perceptions of whether evaluation leads to improvements in their teaching.

FINDINGS

Project Question 1: Is there a difference in teacher perceptions of evaluation in TEAM districts as compared to districts using an alternative observation model?

A. *Teachers report more favorable perceptions of evaluation as a tool for improving teaching in districts using alternative evaluation.*

A greater proportion of teachers in districts using any alternative evaluation (Project Coach, TIGER, TEM) approach view the evaluation process as having led to improvements in their teaching, as compared to teachers in TEAM districts.

B. *Teachers report more favorable overall perceptions of evaluation and feedback in TIGER districts.*

A greater proportion of teachers in TIGER districts view the evaluation process as having led to improvements in their teaching, as compared to TEAM Districts. Additionally, in 2016 and 2017, teachers in TIGER districts were more likely to report receiving feedback from evaluators that prioritizes improving teaching over making judgments about the teacher's performance. This difference did not appear at a statistically significant level in 2018 survey results.

C. *Within TIGER districts, district context is not significantly associated with teacher perceptions of evaluation and feedback – with the possible exception of teacher experience.*

Across TIGER districts, we generally find no statistically significant difference in teacher perceptions of evaluation as a tool for improving teaching when controlling for teacher effectiveness or student demographics. More experienced teachers in TIGER districts may be less likely to report favorable perceptions of evaluation and feedback. In general, teachers across TIGER districts are more likely to report that they receive feedback focused on improving their practice than they are to indicate that they believe evaluation improves their teaching.

Project Question 2: How do district policies and school practices combine to influence patterns of implementation across districts using the TIGER observation model?

A. *TIGER's commitment to flexibility supports continued evolution of the observation model and contributes to varied implementation across contexts.*

While the TIGER observation model includes non-negotiable elements, it also prioritizes implementation flexibility. The flexible options within TIGER produce variation in implementation within and across districts. The combination of non-negotiable state requirements, TIGER non-negotiable core elements, TIGER flexible options, and managerial discretion enable context-specific use of the TIGER observation model.

B. *TIGER's holistic approach to evaluation builds buy-in from teachers and administrators.*

Teachers participating in the TIGER observation model receive a summative evaluation score at the end of the year that serves to comply with state accountability requirements and reflects the sum total of evaluator observations, informal classroom walkthroughs, perceived engagement with professional development and the school community, and other factors. Teachers and administrators collectively appeared to appreciate that the TIGER observation model relies on multiple forms of evidence and provides a holistic, annual score instead of averaging a teacher's observation scores across a year.

C. *While central office leaders evinced a commitment to maintaining a firewall between evaluation and coaching, in practice, that firewall is sometimes porous.*

TIGER calls for robust coaching protocols and a clear separation between evaluation and coaching. Though central office leaders appeared committed to maintaining a firewall between evaluation and coaching, the firewall often appears porous. Teachers perceive that coaches and principals communicate frequently about teachers' progress. However, this does not appear to influence teachers' general perceptions of the TIGER observation model, perhaps due to apparently high degrees of relational trust in the study's target districts.

D. *One of the TIGER observation model's strengths lies in its emphasis on teacher growth, as opposed to compliance or accountability.*

Respondents attributed TIGER's focus on growth to two features of the observation model — its emphasis on feedback and its holistic approach to assigning summative ratings. Commitments to frequent performance feedback, professional development, and public practice permeated the districts we visited and contributed to cultures of improvement. Deemphasizing scores throughout the year seems to help reorient teachers and administrators away from accountability and toward growth.

E. *"Just enough, Just in Time" feedback promotes improvement in TIGER districts.*

The TIGER observation model goes beyond the stated TEAM policy objectives in its efforts to provide targeted, content specific, and timely feedback to drive instructional improvement. Schools we visited worked to deprivatize teaching and to advance ongoing conversations about instructional practice to support growth.

Project Question 3: How do school and district contexts influence perceptions of teacher evaluation in districts using the TIGER observation model?

A. *Contextual factors like district size and quality of relationships may influence perceptions of the TIGER observation model and raise questions about the feasibility of scaling it.*

The small size of TIGER districts may facilitate greater relationship building, trust, and collaboration among faculty and staff, which may have implications for the feasibility of scaling the TIGER observation model to other, larger districts across the state of Tennessee.

B. *Districts in which stakeholders spoke most positively of the TIGER observation model invested in differentiated and ongoing supports for teachers, school leaders, and coaches.*

While the TIGER model prioritizes protected coaching for Stage I teachers, coaching is the most obvious form of support observed within the case study schools. Schools also prioritized regular engagement with professional learning communities (PLCs) and frequent district professional development days. Districts also invest in building the capacity of school leadership and instructional coaches to provide targeted actionable feedback.

C. *Building leaders play a key role in implementing TIGER and establishing cultures of professional growth i their buildings.*

District leaders empowered building principals to adapt TIGER to meet the needs of their faculty. As a result, we observed the variation of TIGER implementation across schools highlighting the key role of building leadership. Principals who focused on leveraging TIGER's flexible options as opposed to complying with evaluation mandates appeared more successful in creating a culture of professional learning in their schools.

RECOMMENDATIONS

Based on these findings, we recommend that TERA and TDOE consider:

Holistic scoring of teacher evaluations. TEAM guidelines currently call for evaluators to determine a teacher's summative score by averaging the teacher's scores across each observation (Hunter 2018). In the TIGER districts we visited, we learned that teachers and administrators valued the opportunity to assign a summative score based on the teacher's performance across the year. This appears to reinforce a growth mindset among school leaders and teachers and built buy-in for the TIGER observation model.

Replacing the generic TEAM rubric with content-specific rubrics. While all TIGER districts use the TEAM general educator rubric for teacher evaluations, we heard from stakeholders in some districts that for informal walkthroughs they prefer using content-specific rubrics from external partners. Our review of research further supports our findings, as it suggests that teachers are more likely to improve in response to content-based, as opposed to generic, feedback (Hill & Grossman, 2013). At minimum, we suggest that Tennessee consider providing TEAM rubrics specific to English Language Arts/social studies and math/science. Ideally, these rubrics also would take into consideration grade bands such as elementary, intermediate, and secondary.

Supporting regular, regional convenings of district leaders to examine and improve teacher evaluation implementation. A compelling feature of the TIGER model is the User Group network. Comprised of district and school leaders, the User Group functions like an improvement community dedicated to cultivating a deep understanding of TIGER implementation, exploring flexible options as implemented in participating districts, and considering adaptations to the observation model (Bryk et al., 2015). We find that this builds buy-in for the observation model and supports successful implementation across districts. We recommend that TDOE consider supporting regional networks of TEAM district leaders focused on examining and improving teacher evaluation in their respective districts, which may strengthen overall implementation of the model and surface ideas for further refinement.

Building mechanisms for non-evaluative coaching into the evaluation system. TEAM provides clear guidance for conducting post-observation conferences in ways that promote teacher reflection and identify clear areas for development (TDOE, 2018, Evaluator Handbook). Less clear is whether TEAM explicitly encourages or requires any sort of coaching for educators. By contrast, the TIGER observation model requires that Stage I teachers receive non-evaluative coaching, and our observations and interviews suggested that many Stage II teachers also take advantage of opportunities to work with coaches. This investment in coaching reflects research suggesting that coaching impacts teachers' instructional improvement (Hill & Grossman, 2013; Kraft, Blazar, & Hogan, in press). Moreover, it reinforces TIGER's commitment to leveraging teacher evaluation for growth and improvement by providing teachers with protected space to develop their skills. We recommend that TEAM incorporate widespread access to protected coaching, at least for early career or less effective teachers.

Incentivizing more public practice and prioritizing timely, actionable feedback within and beyond formal observations. The TIGER observation model deemphasizes formal observations for accountability purposes in favor of encouraging frequent walkthroughs and other, informal opportunities to observe and provide feedback on teacher practice. The prioritization of public practice and informal observation contributed to cultures of growth within many of the buildings we visited. Currently, TEAM allows informal walkthroughs to augment teacher evaluation only for level five educators (TDOE, 2018, Evaluation Handbook). The system might be strengthened and more oriented toward growth if state leaders encouraged the use of frequent walkthroughs for teachers at all levels of effectiveness.

Additionally, stakeholder interviews suggested that when teachers receive frequent feedback from administrators and coaches (and even peers), they feel well-supported and confident in their growth as educators. This suggests a possible role for TDOE, which we encourage to consider creating ongoing professional development aimed at supporting school leaders and instructional coaches in providing content-specific, actionable feedback to teachers. This could be incorporated into existing supports, like the Tennessee Academy for School Leaders, or reflect new offerings geared jointly to administrators and coaches (TDOE, n.d, Tennessee Academy for School Leaders).

INTRODUCTION

In recent decades, research demonstrating both that teacher quality significantly influences student achievement and that teacher effectiveness varies considerably stimulated a number of reforms aimed at improving the productivity of America's teachers (Rivkin, Hanushek, & Kain, 2005; Goldhaber, 2016; Taylor & Tyler, 21012; Howell, 2015; Weisberg, et al., 2009). Spurred in large part by the Obama administration's Race to the Top priorities, states pursued reforms to teacher preparation, recruitment, hiring, retention, compensation, and — most ubiquitously — teacher evaluation (Howell, 2015). In a widely cited study, Kane & Staiger (2012) suggested that schools and districts could more reliably measure and predict teacher effectiveness through multiple measure evaluation ratings that include classroom observations, student surveys, and value-added measures of teachers' impact on student achievement. This research took hold. As of 2015, 27 states require annual evaluations for all teachers, and 43 states mandate that teacher evaluation systems include objective measures of student achievement (NCTQ, 2015).

Tennessee counts itself among the early adopters of teacher evaluation reforms. Tennessee's First to the Top Act (2010) specifies that all teachers must receive annual evaluations, comprised of observations and student achievement and growth data (Offices of Research and Education Accountability, 2012). Currently, the state mandates that between 50-60 percent of a teacher's evaluation rating come from classroom observation, making observation the single most consequential component of a teacher's evaluation (Tennessee State Board of Education, 2017).The majority of Tennessee districts use the state-developed Tennessee Educator Acceleration Model (TEAM) rubric to gather observational data on teacher effectiveness, while a small number of districts use state-approved alternative observation models, either in combination with or separate from the TEAM rubric (Tennessee State Board of Education, 2017; Appendix A).

While recent analyses find little to indicate that teacher evaluation results in meaningful increases in teacher quality, early evidence from Tennessee suggests that the state may buck this trend (Stecher, et al., 2018; Hallinger, Murphy, & Heck, 2013; Taylor & Tyler, 2012; Papay & Laski, 2018). Using teacher observation ratings and value-added scores, Papay & Laski (2018) report that Tennessee teachers improve rapidly during their early career and that they continue to improve as their careers progress. Admittedly, this improvement varies substantially by school and district (Papay & Laski, 2018). Survey data suggests a potential relationship between this improvement and teacher evaluation. Fifty-eight percent of all Tennessee teachers responded to the 2018 Tennessee Educator Survey, which aims to provide policymakers with teachers' perceptions on such topics as instructional time, evaluation, and improvement (TDOE, 2018). The survey's results indicate that teachers perceive evaluation as a tool for improving practice. Seventy-two percent of survey respondents indicated that teacher evaluation improved their teaching (TDOE, 2018). While promising, these various findings raise questions about what elements of the state's teacher evaluation systems influence teachers' perceptions of improvement, and whether these systems operate more effectively in specific types of contexts.

Currently, staff at the Tennessee Department of Education (TDOE) are preparing to review and revise the state's teacher evaluation policies, in the hopes of strengthening the connection between teacher evaluation and teacher improvement. In particular, TDOE leaders have expressed interest in whether features of the state's alternative evaluation approaches may facilitate greater improvement than the more commonly used TEAM evaluation model.

REQUEST FOR ASSISTANCE

In a Request for Assistance to the Vanderbilt Doctor of Education program, the Tennessee Education Research Alliance (TERA) expressed interest in better understanding variation in professional growth and feedback among Tennessee's

department's examination of its teacher evaluation policies by providing the state with analyses that may illuminate promising practices as well as areas for adjustment.

The majority of Tennessee school districts implement TEAM, the state's default evaluation model. However, the state allows districts to apply to implement alternative approaches for classroom observation, and at present, there are four approved alternative models being used in approximately fifteen school districts. These are:

- TIGER - adopted by eleven, small districts across the state, most, but not all, of which are members of the Association of Independent and Municipal Schools (AIMS);
- The Teacher Effectiveness Model (TEM) - implemented by Shelby County, the state's largest district;
- Project COACH - at use in Hamilton County, Tennessee's fourth largest district; and
- The Achievement Framework for Excellent Teaching (AFET) - approved for use in Tennessee's state-run Achievement School District, though it remains unclear whether any schools are using this evaluation approach. [1]

Across alternative models, early analyses from TDOE imply that teachers in TIGER districts more favorably rate teacher evaluation than teachers in TEAM districts (Kevin Schaaf, Tennessee Department of Education, personal communication). TIGER also presents an interesting alternative observation protocol for analysis because its use can be examined across multiple districts contexts (unlike TEM or Project COACH). Given early, positive signs about teacher perceptions and its use across several districts, our project team elected to conduct a deep dive into TIGER districts to explore whether and how features of its observation model and/or conditions in TIGER schools and districts may result in more favorable teacher perceptions.

PROJECT QUESTIONS

In collaboration with TERA and the TDOE, we designed a mixed-methods study that explores the relationship between teacher evaluation models and teachers' perceptions of whether evaluation leads to improvements in their teaching. We aim to support the department's policy review by helping state education leaders better understand:

1. Is there a difference between teacher perceptions of evaluation in TEAM districts as compared to districts using an alternative observation model?
2. How do district policies and school practices combine to create patterns of implementation across districts using the TIGER observation model?
3. How do school and district contexts influence perceptions of teacher evaluation in districts using the TIGER observation model?

We begin our exploration of these research questions by considering the policy context in which Tennessee districts implement teacher evaluation.

[1] Source - alternative evaluation models - Tennessee State Board of Education, 2017; Source - district size - Tatter, 2016.

BACKGROUND & POLICY CONTEXT

In 2010, the TN First to the Top Act passed in conjunction with the federal Race to the Top competition. Tennessee's act specifies that all teachers must receive annual evaluations comprised of observation and student achievement data (TDOE, n.d., Evaluation). Enacted in July 2011, the act specified that all districts in Tennessee implement a state-approved teacher evaluation model (TDOE, n.d., Evaluation). TDOE supported the initial implementation of TEAM across the state during the 2011-12 school year (TDOE, n.d., Evaluation).

Even as state lawmakers were advancing the First to the Top Act, Tennessee's AIMS districts were collaboratively engaged in researching, designing, and developing a teacher evaluation approach, TIGER, which twenty-four AIMS districts piloted during the 2010-2011 school year (AIMS/TIGER Teacher Evaluation Model, 2017). Rather than adopt the TEAM model the following year, several TIGER districts applied for and received state approval to continue the implementation of the TIGER observation model (AIMS/TIGER Teacher Evaluation Model, 2017).

Below, we summarize the primary features of these two approaches to teacher evaluation.

Tennessee Educator Acceleration Model (TEAM)

TEAM remains Tennessee's default evaluation model, with approximately eighty-five percent of districts employing the system (Hunter, 2018). The TEAM observation rubric draws from the widely cited Danielson framework for teacher evaluation and covers four domains: Instruction, Environment, Planning, and Professionalism (Hunter, 2018). Within each domain, teachers receive ratings on multiple indicators (Hunter, 2018). For example, within instruction, teachers receive individual ratings on twelve indicators (Hunter, 2018). Some of these indicators are content neutral (e.g. grouping students) while others require teachers to demonstrate content expertise in addition to pedagogical proficiency (e.g. presenting instructional content) (Hunter, 2018). A teacher's instruction, environment, and planning scores derive from classroom observations, while professionalism ratings come from several sources, including participation in the school community and self-reflection (Hunter, 2018). Teachers receive indicator ratings ranging from one (significantly below expectations) to three (at expectations) to five (significantly above expectations). At the end of each school year, TEAM teachers receive a summative observation indicator rating that reflects the mean of all observation scores across the year (Hunter, 2018).

Tennessee requires that TEAM observers attend training and pass a certification exam developed by the National Institute for Excellence in Teaching, a nonprofit organization that helped state officials develop the TEAM rubric (Hunter, 2018). These policies aim to ensure consistency across observation processes and ratings (Hunter, 2018). A TEAM observation consists of any classroom visit in which teachers receive a score on a single rubric domain (Hunter, 2018). If an observer visits a classroom and provides a rating on, for example, both instruction and environment, then the visit counts as two observations (Hunter, 2018).

Tennessee assigns each teacher to a defined number, duration, and type (announced, unannounced) of observation based on their certification status and effectiveness (as determined by prior evaluation ratings and/or scores on the Tennessee Value Added Assessment System -TVAAS) (Hunter, 2018). The number of required observations range from three to seven, and up to two "observations" – or rating of a domain on a district's observation rubric – can occur in the context of one classroom visit (Hunter, 2018). Tennessee requires a cumulative duration of observations ranging from sixty to ninety minutes across the school year (TDOE, 2018-19 Observation Guidelines, n.d.). The state specifies that at least half of these observations must be unannounced (TDOE, 2018-19 Observation Guidelines, n.d.).

Within the TEAM observation model, observers conduct pre-observation conferences for all announced classroom visits (Hunter, 2018). According to the TDOE, pre-conferences should help observers familiarize themselves with the lesson objectives and content, as well as gather any relevant information about the classroom context, and/or areas of concern for the teacher (TDOE, 2018-19 TEAM Teacher Evaluator Handbook). In his history of TEAM, Hunter (2018) notes that the pre-observation conference can contribute to an observer's ratings in the planning domain. Following any TEAM observation (announced or unannounced), observers should conduct post-conferences with teachers. Tennessee instructs TEAM observers to focus the post-observation conference on just two indicators — an area of relative weakness for the teacher and an area of relative strength (TDOE, 2018-19 TEAM Teacher Evaluator Handbook; Hunter, 2018). Additionally, observers should provide teachers with written feedback within one week of an observation (Hunter, 2018).

Across TEAM, Tennessee officials stress the coaching function of the post-observation conference, providing observers with sample coaching questions and explicitly stating: "It is important to note that a post-conference does not begin with a presentation of the scores, but with coaching questions that, through reflection, lead to the identification of the areas of reinforcement and refinement" (TDOE, 2018-19 TEAM Teacher Evaluator Handbook, 61, emphasis in original). Within the TEAM observation protocol, these observation and feedback cycles are conducted by the individuals who provide ratings on one or more domains for the teacher's evaluation. (However, a teacher may receive observations from different individuals over the course of a year).

Teacher Instructional Growth for Effectiveness and Results (TIGER)

Eleven non-urban, relatively small districts conduct teacher observations using the TIGER protocol (AIMS/TIGER Teacher Evaluation Model, 2017). In 2014-15, 1,765 teachers — or approximately 2.5 percent of Tennessee's teachers — were evaluated using TIGER observation models (Hunter, 2018). AIMS members collaborated in the development of the TIGER observation protocol over the course of 2010 (TIGER: A Formative Model for Teacher Evaluation, n.d.). Since the 2014-15 school year, TIGER districts have used the state's TEAM rubric to conduct formal observations, but the model differs from TEAM in the processes by which teachers are observed, given feedback, and eventually scored (AIMS/TIGER Teacher Evaluation Model, 2017).

According to the AIMS/TIGER Teacher Evaluation Model guidance document (2017), early developers rooted the observation model in a number of principles, chiefly that TIGER:

- emphasizes formative feedback, coaching and professional development, and growth over high-stakes accountability;
- sees teachers as full partners in the evaluation process;
- seeks to "cultivate a trusting, collaborative supervisor-teacher partnership — with formative intent and activity — to the greatest degree possible" (5)
- embraces a differentiated process, such that "teachers with the greatest need for growth occupied the most attention and effort" (5); and
- allows for differentiated compensation as teachers move into different "stages" of effectiveness, with additional performance pay possible for Stage III teachers.

Anchored in these principles the TIGER observation model continues to evolve. Currently, leaders from AIMS districts that use TIGER meet regularly in a forum dubbed the TIGER User Group. This network of school leaders seeks to improve the implementation of TIGER by defining features of the model according to the following categories: coaching stages, TIGER non-negotiable core elements and flexible implementation options, managerial discretion, and state compliance requirements (John Campbell, AIMS TIGER Evaluation System of Tennessee, Board President, personal communication).

Coaching Stages

The TIGER observation model calls for a coaching approach that decreases as teachers move through three defined stages of effectiveness. (See Appendix B for a process map describing how teachers move through stages of effectiveness in TIGER).

<u>Stage I</u>

Teachers receive a Stage I designation if they are new to the profession or their district, or if TDOE assigned the teacher a level of effectiveness rating below three based on the prior year's evaluation.[2] All Stage I teachers receive regular coaching, and TIGER differentiates itself from TEAM in that a teacher's coach is separate from his or her evaluator and theoretically not involved at all in the formal evaluation process (AIMS/TIGER Teacher Evaluation Model, 2017).

At the beginning of the year, the evaluator (most likely the principal) conducts a pre-conference with the Stage I

[2]Regardless of evaluation approach, all Tennessee districts submit teacher summative teacher evaluation ratings to TDOE, which combines teachers' observation and professionalism ratings with measures of student growth to assign an overall level of effectiveness rating ranging from 1 (least effective) to 5 (most effective) (Tennessee Department of Education, TN Report Card, n.d.).

teacher followed by an announced, formal observation (AIMS/TIGER Teacher Evaluation Model, 2017). Following the observation, the teacher, principal, and a non-evaluative instructional coach conference together (AIMS/TIGER Teacher Evaluation Model, 2017). During this conference, the team reflects on the teacher's lesson and identifies areas of strength ("reinforcements" in the Tennessee parlance) and areas for growth ("refinements") (AIMS/TIGER Teacher Evaluation Model, 2017). The coach then begins conducting series of formative walkthroughs and feedback sessions with the teacher (AIMS/TIGER Teacher Evaluation Model, 2017). These conversations remain confidential unless the teacher chooses to share this feedback with the evaluator (AIMS/TIGER Teacher Evaluation Model, 2017). Districts have the option to institute mechanisms whereby the coach can re-engage the evaluator if they determine that the teacher is not making sufficient progress and/or the coaching relationship proves ineffective (AIMS/TIGER Teacher Evaluation Model, 2017). Under typical circumstances, when the coach determines the teacher can justify their practice with evidence, they release the teacher back to the evaluator for a formal unannounced observation (AIMS/TIGER Teacher Evaluation Model, 2017).

Following this unannounced observation, the teacher completes a lesson reflection and engages in a post-conference with their evaluator (AIMS/TIGER Teacher Evaluation Model, 2017). During this conference, the teacher and evaluator can provide additional evidence to support the evaluator's assessment (AIMS/TIGER Teacher Evaluation Model, 2017). Following this conference, the evaluator will review all evidence, confirm the fidelity of the evaluation, and determine a summative observation rating for each indicator on the TEAM rubric (AIMS/TIGER Teacher Evaluation Model, 2017). In TIGER, evidence for ratings can come from a variety of sources including formal observations, informal observations, and other indicators of the teacher's competency in a rubric domain (AIMS/TIGER Teacher Evaluation Model, 2017). This information is shared with the teacher, along with suggested areas of refinement and reinforcement, and submitted to TDOE, which then determines the teacher's overall level of effectiveness based on evaluation ratings and measures of student growth and achievement (AIMS/TIGER Teacher Evaluation Model, 2017).

Stage II

Stage II teachers are experienced teachers with an effectiveness rating equal to or greater than 3 (AIMS/TIGER Teacher Evaluation Model, 2017). At the start of the year, the teacher and evaluator complete a collaborative Reflection and Growth Plan (AIMS/TIGER Teacher Evaluation Model, 2017). The Reflection and Growth Plan emphasizes teacher self-assessment and identifies areas for reinforcement and refinement, goals for professional growth, and strategies for professional learning during the year.

The plan determines the focus for evaluator walkthroughs and feedback sessions, which should occur at least twice a semester. At least half of these walkthroughs must be unannounced (AIMS/TIGER Teacher Evaluation Model, 2017). They must last a minimum of 15 minutes and cumulatively meet state requirements for observation time based on a teacher's licensed status (AIMS/TIGER Teacher Evaluation Model, 2017). Evaluators can score walkthroughs against the TEAM rubric if they choose (AIMS/TIGER Teacher Evaluation Model, 2017). Regardless of whether they score

walkthroughs, evaluators must provide qualitative feedback to teachers within five days of a walkthrough (AIMS/TIGER Teacher Evaluation Model, 2017). Mid-year, teachers return to the Reflection and Growth plan and score themselves, providing evidence to justify their rating (AIMS/TIGER Teacher Evaluation Model, 2017).

At the end of the school year, the evaluator reviews evidence of the teacher's effectiveness, including walkthrough and optional formal observation data, feedback and dialogue, reflection and growth plan data, student work, and teacher professional learning (AIMS/TIGER Teacher Evaluation Model, 2017). The evaluator develops a summative rating for the teacher across the domains of the TEAM rubric, calculates an overall summative rating, and identifies areas of reinforcement and refinement for the coming year (AIMS/TIGER Teacher Evaluation Model, 2017). After conferring with the teacher, the evaluator submits summative ratings to TDOE, which calculates their overall level of effectiveness (AIMS/TIGER Teacher Evaluation Model, 2017).

Stage III

TIGER also includes the Stage III designation for highly effective teachers. Stage III teachers follow all of the evaluation procedures detailed in Stage II (AIMS/TIGER Teacher Evaluation Model, 2017), however, Stage III teachers may assume additional leadership roles within a district, such as coaching, leading PLCs, or serving as a peer evaluator (AIMS/TIGER Teacher Evaluation Model, 2017). Individual TIGER districts determine what, if any, Stage III teacher leader opportunities to make available (AIMS/TIGER Teacher Evaluation Model, 2017). Stage II teachers are eligible to apply for Stage III status. Districts develop their own application processes and eligibility criteria for Stage III designation (AIMS/TIGER Teacher Evaluation Model, 2017). Additionally, districts may consider providing strategic compensation for Stage III teachers (AIMS/TIGER Teacher Evaluation Model, 2017).

Non-Negotiable Core Elements & Flexible Implementation Options

All TIGER districts agree to adhere to a set of core elements of the observation model, specifically:

- Use of the TEAM observation rubric;
- Relying on summative scoring rather than an average of observation ratings across the year;
- Providing a differentiated evaluation process based on (at minimum) Stages I and II;
- Providing protected, non-evaluative coaching for Stage I teachers;
- Promoting collaborative reflection; and
- Using multiple forms of evidence to arrive at a holistic, summative rating.[3]

We wish to call particular attention to TIGER's reliance on holistic, summative scoring. In this, TIGER differs importantly from TEAM, which averages a teacher's scores across multiple observations to arrive at a teacher's summative rating (Hunter, 2018). Proponents of TIGER's 'no averaging' rule believe that it fosters a growth mindset and reinforces the formative feedback and coaching supports called for in the model (TIGER: A Formative Model for Teacher Evaluation, n.d.).

Within all these non-negotiables, TIGER provides considerable implementation flexibility to districts. Table I highlights key features of TIGER's non-negotiable core attributes and related flexibilities (TIGER User Group, 2018).

[3] (TIGER User Group, 2018)

Table I

	TIGER Non-Negotiable Core Elements	TIGER Flexible Implementation Options
State Rubric	TIGER utilizes the TEAM rubric and TDOE TEAM observer training.	Districts may use alternative research-based criteria, aligned with the TEAM rubric, for meeting the TEAM rubric's indicators, to meet district or school objectives (e.g., School Achievement Partners Instructional Planning Guides).
No Averaging	Interim observations or walk-through ratings may not be averaged to calculate year-end indicators summative scores.	Districts may choose to score interim walk-through observations, though only qualitative feedback is required. Observation feedback does not have to include every indicator.
Differentiated Process	Implementation is differentiated into three different processes, called "stages," depending on teacher need.	Districts may set their own cut-points for stage determinations. Districts may vary Stage III processes, based on resources and needs. Districts may vary time periods for some steps in the annual process.
Protected Coaching	Stage I teachers are observed and receive feedback from non-evaluating coaches in a protected, formative process.	Coaching options include Stage III teachers serving as peer-coaches. Coaching and Stage I processes may be used with any teacher, regardless of stage.
Collaborative Reflection	Teachers and principals collaboratively reflect and negotiate the relationships between evidence and reflective and summative ratings.	Teachers' annual, required, self-reflection activity and documentation may occur at any time during the year, at each district's discretion.
Holistic Evaluation	The evaluation process is driven by authentic feedback, collaborative evidence collection, and evaluative credit for growth during the year.	Evidence may be collected at any time — within or outside of specific observation events — by teachers as well as observers. Districts may enhance the process to increase the effectiveness of evaluation for teacher growth.

Managerial Discretion

Flexible management options allow for districts to adapt the implementation process to their district context. Examples of exercising managerial flexibility include the adoption of technology platforms to manage the observation model; the adaptation of observation, reflection, and feedback forms; or, the timing of prescribed teacher reflections and activities throughout the school year.

State Compliance

TIGER requires participating districts to comply with state policy by ensuring that 50% or more of classroom observations are unannounced; that every teacher has multiple observations; that experienced professionally licensed teachers are observed for a minimum of sixty minutes per year; apprentice teachers are observed for a minimum of ninety minutes per year; and that teachers receive feedback within five days of observation (TSBE policy 5.201).

CONCEPTUAL FRAMEWORK & LITERATURE REVIEW

Conceptual Framework

Because our research questions are best understood within the broader frame of teacher evaluation as a lever for instructional improvement, our conceptual framework seeks to illuminate the conditions and processes that may support instructional improvement. Given our policy context and research questions, we first examine how teacher evaluation may contribute to instructional improvement. Then, we look to the extant literature related to *school context* — especially as it pertains to administrative support and teacher collaboration — to understand the conditions under which teacher evaluation will most likely support teacher instructional improvement. Finally, we explore the specific mechanisms by which teacher evaluation systems might facilitate instructional improvement, namely via *observation, feedback, and coaching*.

Literature Review

Teacher evaluation & teacher improvement

Proponents of teacher evaluation cite two mechanisms by which it may increase teacher quality. Some see evaluation's primary value in enabling the dismissal of ineffective teachers and retention of effective ones (Papay & Richard, 2018; Papay, 2012; Taylor & Tyler, 2012; Hill & Grossman, 2013; Stecher, et al., 2018; Kraft & Gilmour, 2016). This perspective aligns with the notion that, following a period of rapid improvement in their early careers, teachers' effectiveness largely stagnates (Papay & Laski, 2018; Papay & Kraft, 2015; Kraft & Papay, 2014; Taylor & Tyler, 2012). Emerging analyses seem to refute both these hypotheses. One body of research suggests that teachers indeed improve throughout their careers (Kraft & Papay, 2014; Papay & Kraft, 2017; Taylor & Tyler, 2012; Papay & Laski, 2018). Another demonstrates that the vast majority of teachers continue to receive high ratings under new evaluation systems, suggesting that evaluation processes rarely facilitate the dismissal of ineffective teachers (Taylor & Tyler, 2012; Weisberg, et al., 2009; NCTQ, 2015; Grissom & Loeb, 2017; Papay & Richard, 2018; Hill & Grossman, 2013; Stecher, et al., 2018). Critics of this view of teacher evaluation assert that widespread adoption of teacher evaluation reforms has failed to yield demonstrable improvements in overall teacher quality (Hallinger, Heck, & Murphy, 2013; Stecher, et al., 2018).

Another view of teacher evaluation, more aligned with Tennessee's goals, holds that effective teacher evaluation systems can promote long-term improvement in teacher performance by providing teachers with meaningful information about the quality of their practice (Papay, 2012; Hill & Grossman, 2013; Papay & Richard, 2018; Taylor & Tyler, 2012; Sun, Mutcheson, & Kim, 2015; Kraft & Gilmour, 2016). In the past decade, a number of studies have suggested that participation in robust teacher evaluation systems can increase teacher effectiveness. Analyzing a sample of mid-career math teachers who participated in a year-long, observation-based evaluation in the Cleveland Public Schools, Taylor & Tyler (2012) employ a fixed-effects approach to demonstrate that teachers improved their effectiveness during the year they participated in the evaluation and in subsequent years. Dee & Wyckoff (2015) evaluated the early implementation of Washington, DC's multi-measure IMPACT teacher evaluation system, which included both powerful sticks (the threat of dismissal for ineffective teachers) and carrots (sizable performance incentives for highly effective teachers). They present regression discontinuity estimates indicating that teachers with poor evaluation ratings voluntarily left the district at greater rates than otherwise would be expected and that teachers with higher ratings responded to financial incentives by increasing productivity (Dee & Wyckoff, 2015). Evaluating a pilot teacher evaluation system randomly implemented in some Chicago schools, Steinberg & Sartain (2015) find positive effects on reading performance in schools participating in the observation-based system.

These studies merit attention for several reasons. First, they include both veteran and early career teachers, reinforcing the notion that teacher effectiveness remains malleable well into a teacher's tenure. Reviewing data from Tennessee, Papay & Laski (2018) find that although the improvement trajectory appears steepest during a teacher's first three years in the classroom, teachers continue to demonstrate improvement over the course of their careers. Using estimates of within-teacher returns to experience, Kraft, Papay, & Chi (2018) find substantial rates of improvement across a teacher's first ten years on the job. Kraft & Papay (2014) report similar findings, though they suggest that school context may mediate this improvement trajectory (see below).

Additionally, the studies described above featured particularly robust evaluation systems, characterized by a high degree of feedback and support. Indeed, Taylor & Tyler (2012) acknowledge that the effects they report may stem from the fact that teachers in their study had not received detailed feedback on their performance in years, and that teachers who more regularly receive such feedback may not respond as dramatically to the introduction of evaluation. Dee & Wyckoff (2015) carefully note that DC leaders designed IMPACT such that "the expectations of teachers were clearly articulated and communicated and teaching support to meet those expectations (e.g. instructional coaches) was available" (293). Similarly, Steinberg & Sartain (2015) emphasize that Chicago leaders decoupled the pilot evaluation program from accountability frameworks and, especially in the first year of the program, provided intensive supports to principals around observation, feedback, and coaching. Perhaps as a result of these lower stakes and more intensive supports, Steinberg & Sartain (2015) find more substantial effects for the program among schools participating in the first pilot year.

In point of fact, other studies imply that less robust or formatively oriented evaluation systems may not produce similar improvements in teacher practice. For example, Stecher, et al. (2018) reviewed the wide-scale implementation of teacher evaluation reforms supported by the Bill & Melinda Gates Foundation in a number of urban districts. Their analyses find minimal evidence that evaluation promoted increases in teacher quality, but also note that many participating districts sought to mitigate time burdens on evaluators by reducing the length and/or frequency of evaluations and that districts struggled to align development opportunities to teachers' individual evaluations (Stecher, et al., 2018). Somewhat similarly, Koedel, et al. (2018) report that typical teachers in Tennessee do not use

summative evaluation to inform their professional development, while simultaneously acknowledging that teachers may respond more readily to more formative, ongoing feedback. Therefore, these apparently contradictory findings may actually buttress the notion that the effectiveness of an evaluation system depends on the robustness of its design and implementation. The degree to which teachers experience a robust evaluation system, and therefore increase their likelihood of improving, correlates closely with the school context in which they work and the extent to which the evaluation incorporates high-quality observation, feedback, and coaching.

Creating Conditions for Teacher Improvement: School Context

Significant bodies of research detail the importance of school context in facilitating student outcomes and retaining teachers (Kraft & Papay, 2014; Papay & Kraft, 2017; Bryk, et al., 2010; Kyndt, et al., 2016; Murphy & Torre, 2014; Ingersoll, 2001; Boyd, et al., 2011; Ronfeldt, 2012; Loeb, Hammond, & Luczak, 2005). Kraft & Papay (2014) find that when teachers work in more supportive school contexts — as defined by frequent collaboration, ample opportunities to receive meaningful feedback, and cultures of recognition for strong work — they improve at greater rates than teachers working in less supportive contexts. Using administrative and survey data, including student achievement information, Kraft & Papay (2014) estimate that over a ten-year period teachers working in schools with professional environment ratings in the top quartile improve thirty-eight percent more than teachers working in bottom quartile schools. These findings align with Bryk, et al.'s (2010) longitudinal analyses showing that schools strong in three of five essential supports (school leadership, professional capacity, school-community ties, student-centered learning climate, and instructional guidance) are significantly more likely to improve students' attendance, reading, and math scores. Among the school conditions most relevant to teacher evaluation are teacher collaboration and administrative support. The strength of these conditions may influence the extent to which teachers receive and are able to effectively respond to meaningful feedback about their instruction.

Teacher Collaboration

Collaborative structures — such as opportunities to observe instruction, engage in reflective conversation about instruction, and/or work together to develop curriculum, analyze data, or plan instruction — are associated with greater teacher productivity and student learning gains (Bryk, et al., 2010). Murphy & Torre (2014) posit that cultures of collaboration build teachers' content knowledge and pedagogical skills, while also creating constructive social bonds within schools. Recent empirical research reinforces these perspectives. Jackson & Kirabo (2009) leverage longitudinal data from North Carolina to demonstrate that a teacher's students post larger achievement gains when s/he works with more effective colleagues, suggesting that some degree of peer learning occurs in these school environments. Analyzing Tennessee's Instructional Partnership Initiative, which paired low-evaluated teachers with high-performing peers for the purposes of collaborating around teaching skills, Papay, et al. (2015) find that students in participating schools scored higher than students in control schools, and that gains appeared concentrated among teachers targeted for participation in the initiative. Similarly, Sun, Loeb, & Grissom (2016) report that when less effective teachers move to schools with more effective colleagues, their performance improves. Though this particular study did not extend to the mechanisms through which that improvement occurs, it speaks to the likelihood that teachers exert influence over each other's instructional practices (Sun, Loeb, & Grissom, 2016). In descriptive analyses, teachers report that engaging in well-organized instructional teams improved their classroom practice (Papay & Kraft, 2017). Taken together, these findings suggest that collaboration among peers, especially when intentionally structured around instructional improvement, may contribute to increased teacher effectiveness.

Administrative Support

The quality of school leadership and administrative support represents perhaps the most important factor in successful implementation of a teacher evaluation system. Administrators largely create the conditions for effective

teacher collaboration, feedback, and coaching (Bryk, et al., 2010; Papay & Kraft, 2017; Kraft & Gilmour, 2016). Citing leadership as the driver for school improvement, Bryk, et al. (2010) note that successful school leaders simultaneously attend to the core of teaching and learning by ensuring coherent curricular and instructional programming while developing collaborative structures and cultures of trust and feedback to support teacher growth. "[S]chool leaders," they write, "must nurture over time a new norm formation where continuous individual and collective efforts at improvement within a shared instructional guidance system are 'what we do here'" (206). Murphy & Torre (2014) echo this sentiment, viewing the principal as the "prime actor" in creating productive school cultures (44). Furthermore, empirical studies link principal leadership to improved student outcomes (Bryk, et al., 2010; Supovitz, Sirinides, & May, 2010).

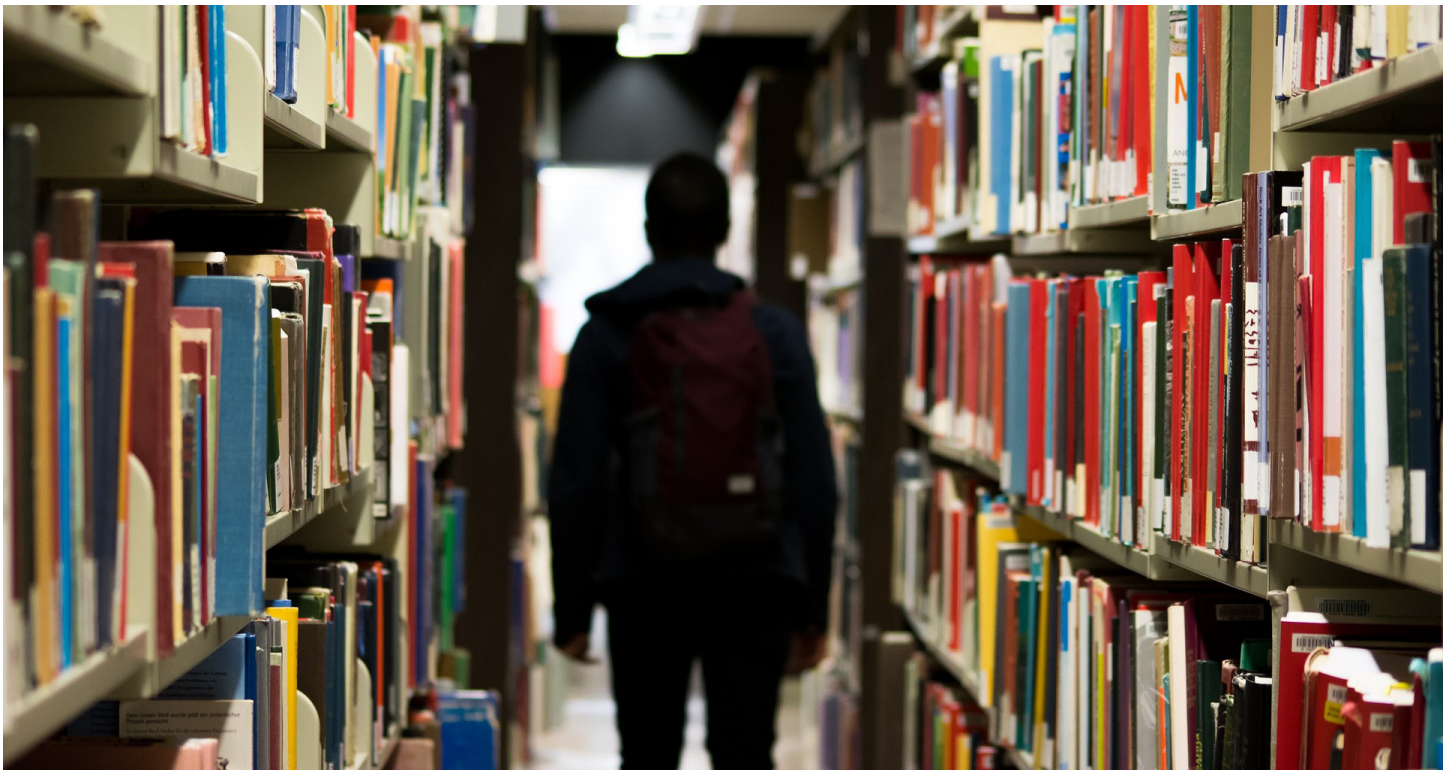
While Neumerski, et al. (2018) observe that principals typically receive only vague guidance about how to enact these crucial instructional leadership roles, some researchers do find that when principals dedicate time specifically to coaching and evaluation student achievement gains ensue (Grissom, Loeb & Master, 2013). This suggests that teacher evaluation may function as a mechanism for strengthening administrative support and teacher improvement. Indeed, school administrators serve as a linchpin in most teacher evaluation systems, which typically place a heavy emphasis on administrators' classroom observations (Papay, 2012; NCTQ, 2015; Campbell & Ronfeldt, 2018; Neumerski, et al., 2018). In the following section, we explore observation as a mechanism for teacher improvement and how well principals facilitate its use as such. Here, we consider more generally principals' functional role in teacher evaluation systems.

Kraft & Gilmour (2016) observe that as primary evaluators in most systems, principals face increased demands on their time and higher expectations for their capacity to serve as instructional leaders. Ideally, principals would meet this expectation by faithfully conducting observations, providing specific feedback to teachers, and facilitating ongoing faculty collaboration around instruction (Kraft & Gilmour, 2016; Donaldson, 2013). In reality, principals have historically spent relatively little time focused on instructional activities and often find it challenging to reorient their time to support evaluation (Kraft & Gilmour, 2016; Grissom, Loeb, & Master, 2013; Donaldson, 2013). Donaldson (2013) and Stecher, et al., (2018) reinforce this perspective, noting that principals complain they lack time to implement evaluation with fidelity. In a recent evaluation, Stecher, et al. (2018) report that many schools and districts respond to these laments by constraining the evaluation system (e.g. reducing the number of required observations) rather than by meaningfully restructuring principals' duties. In other instances, it seems that districts do not accompany teacher evaluation mandates with robust training and support to ensure that principals understand how to complete effective observation and feedback cycles (Neumerski, et al., 2018; Grissom, et al., 2017; Steinberg & Sartain, 2015). Finally, principals in some qualitative studies perceive declines in teacher-principal trust in the context of evaluation, which may stem from teacher concerns about the accountability aspects of evaluation or whether or not the principal possesses sufficient content expertise to accurately assess their instruction (Kraft & Gilmour, 2016; Neumerski, et al., 2018).

Despite these frustrating findings, the literature implies that, if well-implemented and supported, teacher evaluation systems could facilitate strong instructional leadership on the part of school leaders. For example, observation rubrics may illuminate elements of good instruction for principals and provide them with a common language for discussing instruction with faculty (Neumerski, et al., 2018; Kraft & Gilmour, 2016). Moreover, the process of collecting evidence to satisfy evaluation requirements theoretically builds principals' knowledge of their teachers' instructional capacity and facilitates more concrete feedback (Papay, 2012; Neumerski, et al., 2018; Kraft & Gilmour, 2016). Indeed, qualitative analyses conducted by Kraft & Gilmour (2016) and Nuemerski, et al. (2018) indicate that despite the challenges posed by teacher evaluation systems, principals report that implementing teacher evaluation increases their understanding of and focus on strong instruction. In numerous analyses, evaluation expectations around observation and feedback emerge as the primary driver behind principals' enhanced instructional roles (Neumerski, et al., 2018; Grissom, et al., 2013; Kraft & Gilmour, 2016). These findings point to the importance of observation and subsequent feedback and coaching as mechanisms for teacher improvement.

Mechanisms for Teacher Improvement: Observation, Feedback, and Coaching

If school context represents the external conditions under which evaluation systems are most likely to be well implemented, then the quality of an evaluation system's observation and feedback cycles represent the internal mechanisms most likely to facilitate improvement. Indeed, feedback appears to drive Tennessee teachers' assessment of the usefulness of teacher evaluation. In 2017, teachers who rated the evaluation system as helpful for improving practice were more likely to report both receiving feedback in general and receiving targeted feedback on their strengths and weaknesses than teachers who rated the evaluation system less positively (TDOE, 2017).



Observation

As previously noted, observation of classroom practice has emerged as the key feature of most states' approaches to teacher evaluation (Papay, 2012; NCTQ, 2015; Campbell & Ronfeldt, 2018; Kraft & Gilmour, 2016; Neumerski, et al., 2018). To be precise, Papay & Richard (2018) report that forty-six states currently require observation as a component of teacher evaluation. Some point to limitations in the use of administrator observations, pointing to limited variability in principals' overall ratings of teachers, which suggests that observations do little to differentiate high-performing from lower-performing faculty (Weisberg, et al., 2009; Grissom & Loeb, 2017; Papay, 2012; Papay & Richard, 2018). Other scholars raise important concerns about equity, noting that teachers of color and teachers serving larger proportions of students of color, low-income, or historically low achieving students receive, on average, lower observation ratings than white peers or peers teaching more advantaged students (Campbell & Ronfeldt, 2018). These lower ratings do not appear to correlate with teachers' value-added contributions to student learning (Campbell & Ronfeldt, 2018).

Despite these concerns, empirical analyses suggest that administrator observations serve as valid and reliable measures of classroom practice (Cohen & Goldhaber, 2016; Campbell & Ronfeldt, 2018; Kraft, Papay & Chi, 2018; Grissom & Loeb, 2017; Jacob & Lefgren, 2008). Jacob & Lefgren (2008) asked elementary principals in a midsize, Midwestern district to rate teachers along a variety of performance dimensions. They then compared these ratings to value-added estimates of the teachers' effectiveness, finding that principals generally can differentiate between

low- and high-performing teachers (Jacob & Lefgren, 2008). Grissom & Loeb (2017) addressed head on the question of limited variability in principals' performance ratings on teacher evaluations. They interviewed principals in a large, urban district, asking them to rate the effectiveness of a sample of teachers (Grissom & Loeb, 2017). Grissom & Loeb (2017) then compared those low-stakes ratings with the high-stakes evaluation ratings the principals gave the same teachers, finding much more variability in principals' low-stakes ratings. This simultaneously suggests that principals in fact can differentiate among teachers' performance and that they demur from doing so in high-stakes settings. This presents a challenge for policymakers seeking to use high-stakes evaluation as a lever to drive instructional improvement (or dismissal decisions). Nevertheless, Papay & Richard (2018) report that teachers in Tennessee who receive more frequent observations and who teach in schools where observers substantially differentiate performance ratings do appear to improve at greater rates than teachers who do not experience these evaluation features, potentially lending credence to the use of observation to improve performance.

Administrators require a great deal of support to effectively deploy observations as a means of differentiating teacher performance and providing concrete feedback. As noted, principals encounter a number of barriers to implementing observations and using teacher effectiveness data (Kraft & Gilmour, 2016; Donaldson, 2013; Grissom, et al., 2017). In districts that successfully ameliorate those barriers by supporting principals in conducting observations, analyzing data, and engaging in instructional conversations with teachers, school leaders more readily employ teacher effectiveness measures in talent management decisions (Grissom, et al., 2017). Steinberg & Sartain (2015) observed more instructional improvement during the first year implementing a pilot evaluation system, which they connect to diminished central office supports to principals in the pilot's second year. Taken together, extant literature suggests that observation, while imperfect, nevertheless functions as a valid and reliable means of gauging teacher performance, and that principals, when properly supported, can use observations to differentiate between levels of effectiveness and spur instructional improvement. To achieve this latter aim, administrators must couple observations with high-quality feedback and coaching.

Feedback and Coaching

A growing body of literature examines the relationship between observation, feedback, and instructional improvement. While feedback and coaching need not occur in the context of a teacher evaluation system, the expectations for observation and feedback within teacher evaluation systems aligns closely with longstanding definitions of coaching (Kraft, Blazar, & Hogan, in press). In terms of feedback, evidence suggests that frequency and specificity matter most for driving instructional improvement. Papay & Richard (2018) report that teachers in Tennessee who receive more frequent observations and feedback improve more rapidly than teachers who do not. Koedel, et al. (2018) conjecture that while teachers only minimally respond to summative feedback, they may respond more readily to consistent, formative feedback. Hill & Grossman (2013) note that teachers improve most when they receive highly individualized feedback, which includes "specific, actionable items they can implement during their work with students" (Hill & Grossman, 2013, 379). This requires that observers drill down into what typically are fairly broad evaluation rubrics to provide feedback at an appropriate grain size to teachers (Hill & Grossman, 2013).

Reviewing coaching literature, Hill & Grossman (2013) report that feedback and coaching following observations can positively affect student outcomes. Kraft & Blazar's (2017) evaluation of a coaching initiative implemented in New Orleans charter schools reinforced this notion, finding that coached teachers score .59 standard deviations higher on an index of effective instructional practices than their non-coached peers. Hill & Grossman (2013) also posit that observers must accompany their feedback with follow-up to ensure teachers make progress against improvement goals and to address any new challenges teachers may experience. This aligns with Grissom, Masters, & Loeb's (2013) observation that the more time principals spent coaching teachers, the better schools in their study performed in both math achievement and growth.

Kraft, Blazar, & Hogan (in press) recently conducted a meta-analysis of sixty experimental and quasi-experimental studies to examine the effect size of coaching on teaching practice. They find large, positive effects of coaching on instructional improvement (Kraft, Blazar, & Hogan, in press). While acknowledging considerable variation in effect sizes across programs, Kraft, Blazar, & Hogan's (in press) analysis finds a pooled effect size of .49 standard deviations, suggesting that effective coaching indeed may improve instructional practice. The authors find smaller, but still positive, effect sizes when they examine the relationship between coaching and student achievement as measured by state standardized tests (.12 SD) (Kraft, Blazar & Hogan, in press). The same study did not find significant differences in effect size among types of schools (elementary, middle, or high), nor did the authors' analysis suggest that the quantity (or dosage) of coaching significantly influenced its effect size (Kraft, Blazar, & Hogan, in press). These findings merit attention for two reasons. First, they suggest that observation and feedback may improve teacher practice across the K-12 continuum. Additionally, they indicate that the quality of observation and feedback likely matters more than the quantity, implying that evaluators can meaningfully influence instructional practice even within the constraints of an evaluation system.

While administrators most frequently provide feedback in the context of evaluation systems, it is important to note that they need not always be the one to assume a coaching role. As previously discussed, an initiative in Tennessee to pair lower-rated teachers with higher-rated peers resulted in improvements in the lower-rated teachers' instruction (without harming that of the higher-rated peer) (Papay, et al., 2015). Given that a coach's content knowledge appears to influence their ability to support teacher improvement, the ability to diffuse coaching responsibilities beyond an administrator and among content experts could contribute to teachers' instructional improvement in response to evaluation feedback (Stecher, et al., 2018; Hill & Grossman, 2013). Overall, our extant literature review suggests that robust teacher evaluation systems, implemented in strong organizational contexts, may support teachers' instructional improvement.

DATA & METHODS

Overview

We employed a mixed-methods analysis, first conducting a quantitative analysis of statewide teacher survey data and teacher value-added data provided by TERA in partnership with TDOE. These statewide data sets informed our understanding of our first research question: Is there a difference between teacher perceptions of evaluation in TEAM districts as compared to districts using an alternative evaluation model? We augmented these empirical analyses with qualitative examinations of four districts using the TIGER observation model. Our qualitative analyses targeted our second and third research questions, which seek to understand how school and district policies, practices, and context create patterns of implementation and influence teacher perceptions of the TIGER observation model.

Quantitative Analyses

Data

Our quantitative analysis addresses our first research question — the degree to which there is relationship between teacher perceptions of evaluation in TEAM districts and districts using alternative evaluation models, and specifically the TIGER model. Using data from the Tennessee Educator Survey Data, we explored the relationship between teacher perceptions of improvement and alternative observation models. We conducted a series of analyses using two items on the survey:

- "In general, the teacher evaluation process used in my school has led to improvements in my teaching." This is an item to which teachers respond on a Likert scale that includes the following options: strongly disagree,

disagree, agree, or strongly agree.

- "During the current school year, the feedback I received from my evaluator was …" Teachers respond to this item by selecting one of the following responses:

  –"Focused more on helping me improve my teaching than making a judgment about my performance."
  –"Equally focused on helping me improve my teaching and making a judgment about my performance."
  –"Focused more on making a judgment about my performance than helping me improve my teaching."

Our respective analyses concentrated on what we deemed favorable responses to these two questions, namely the proportion of teachers who either "agree" or "strongly agree" with the first item and who select "focused more on helping me improve my teaching …" on the second item.

In order to understand the relationship between teachers in TEAM districts, alternative evaluation districts, and TIGER districts, we used teacher level data from the TN Educator survey between 2015-2018. This survey is designed and hosted by TERA, which provided us the data, and administered by TDOE. We begin our analyses with the 2015 results because this represents the first year that the state achieved a response rate greater than fifty percent.

Table 2.

TN Educator Survey Sample

	2014-2015	2015-2016	2016-2017	2017-2018
TEAM/TAP[4]	56,549	58,474	56,579	59,347
Alternative	13,680	11,708	12,199	12,289
TIGER	1,824	1,655	1,637	1,732
TIGER Target Districts	479	520	478	513
N	70,299	70,432	68,778	71,636

To complement the survey response data, we merged it with teacher level Tennessee Value-Added Assessment System (TVASS) estimates using a de-identified linking variable for the 2016-2017 and 2017-2018 administrative years. We used teacher TVASS scores to control for teacher effectiveness in the two regression models we describe in the next section. Because our data linked teacher responses to the district in which they teach, we were able to create district-level variables to support our analyses.

[4] TDOE and the National Institute for Excellence in Teaching (NIET) collaboratively designed the TEAM observation rubric. NEIT's TAP rubric is the precursor to the TEAM rubric (Hunter, 2018). The TAP system is still implemented in a small number of Tennessee districts.

Methods

Z-Test for the Difference in Proportions

Our first analysis sought to understand whether teachers in TEAM districts and teachers in districts using any of the state's alternative evaluations report differing perceptions of teacher evaluation as a tool for improvement or whether the feedback they receive was oriented toward improvement as opposed to judgment. To do so, we conducted a z-test for the difference in proportions of educators responding favorably to the identified items on the Tennessee Educator survey from TEAM districts and alternative districts in each year (2015, 2016, 2017, 2018).

Given our particular interest in TIGER, we then repeated the above analysis but created a new variable for TIGER districts and conducted the z-test for the difference in proportions to compare responses from TEAM districts to those from TIGER districts.

Ordinary Least Squares Regression

We next conducted a series of Ordinary Least Square regressions to examine teacher responses after controlling for factors like years of teacher experience, teacher effectiveness (TVASS) level, and district demographics (as measured by the proportion of students eligible for Free and Reduced-Price Meals (FARMS). For this latter variable, we analyzed teacher responses based on the proportion of FARMS eligible students they teach. The state provides this data sorted into quartiles.

The first regression analyses used TIGER districts as the independent variable each of the two survey items as the dependent variable. For the second survey item, we replaced the categorical variable with a binary variable indicating the favorable response, "During the current school year, the feedback I received from my evaluator was focused more on helping me improve my teaching than making a judgment about my performance." We then repeated this analysis using only the four target TIGER districts as the independent variable and the two survey items as the dependent variable.

Because our quantitative analyses focus on a subset of four TIGER districts (described more fully in our qualitative methods section), we wanted to understand whether any significant differences emerged in the patterns of responses from those specific districts. Therefore, we repeated our regression analyses using our target districts as the independent variable.

Analysis of Variance within Target Districts

To further understand potential differences in patterns of response across our four target TIGER districts, we conducted a repeated cross-section of ANOVAs comparing teacher responses to the identified survey items within our target districts between 2015-2018. When the results yielded significant F-statistics, we then conducted a *post-hoc* Scheffe Test to determine which pairs of means between districts were significant. Again, we replaced the second survey item with a binary variable indicating the favorable response.

Qualitative Analyses

Our qualitative approach aimed to describe the district contexts and patterns of implementation that appear to influence participants' perceptions of the TIGER observation model via semi-structured qualitative interviews and artifact analysis (Patton, 2002). Ideally, rich qualitative descriptions of phenomena will generate insights that inform readers' understanding of the processes, relationships, situations, systems, and people whose interactions create the phenomena in question (Peshkin, 1993). As such, qualitative research allows for a level of in-depth study that, while somewhat limited in its generalizability, nevertheless extends our understanding of concepts and illuminates broader themes that may influence future policy and practice (Patton, 2002; Peshkin, 1993; Shulman, 1981).

Data & Sample

We purposively selected four TIGER districts for our qualitative analyses. We targeted these districts because, of Tennessee's TIGER districts, these four had sufficient teacher response rates on the 2018 Tennessee Educator Survey to generate a district-specific report. This combination of data sources allowed us to connect themes from our qualitative interviews with quantitative analyses of district survey responses. After conducting email and phone outreach to share the purpose of our study, we secured written permission from each district's Director of Schools (i.e. superintendent) to participate in the study. Because districts, and district staff, participated voluntarily in our study, we made a number of commitments to encourage participation. First, we proposed using pseudonyms for participating districts. While we cautioned leaders that our selection criteria (educator survey responses) and the relatively small number of TIGER districts statewide may render total anonymity impossible, the research team agreed to anonymize participating districts to the greatest extent possible. (For this reason, we do not list the geography of each district other than to say here the four participating districts are located in east, middle, and west Tennessee). Additionally, we committed to anonymizing staff responses and to only reporting responses in aggregate, given the relatively small number of staff in each district who participated in our study. Accordingly, when we present themes in our findings, they represent concepts that resonated across all participating districts. In the case of outlier responses that bear mentioning, we take extra steps not to name the district in question.

Below, we provide a summary overview of participating districts and staff. Then, we provide additional context about each participating site before describing our methods of data collection and analysis.

<u>Participating Districts</u>

The Tennessee Department of Education's TN Report Card (n.d.) data provides the following information about participating districts during the 2017-18 school year:

Table 3.

	District A	District B	District C	District D	MEAN
Total Students	1,994	2,228	1,232	1,268	1,681
% Black, Hispanic, or Native Students	34.1	34.4	19.4	36.3	31.1
% White Students	63.6	64.6	80.4	63.4	68
% Economically Disadvantaged Students	25.6	32	36.3	49.9	35.95
% Language Learners	3.4	11.5	1.6	2.1	4.7
% Students with Disabilities	12.4	13	14.9	13.4	13.4
% Proficient in Math	35.3	35	41.1	27.8	34.8
% Proficient in ELA	38.9	31.3	42	28.8	35.3
Composite TVAAS Level[5]	5	5	4	2	4
Number of Teachers	116	136	87	79	104.5
% Teachers Retained one Year to Next	94.8	86.8	88.8	83.1	88.4

While none of the districts in our study serve major metropolitan areas, they vary in their rurality. In a 2016 ranking of Tennessee counties from most rural to most urban, one participating district is located in a county designated among the 25 most rural, and three participating districts are located in the 25 least rural counties (Roehrich-Patrick, Moreo, & Gibson, 2016). Additionally, each participating district is relatively small, with between three and four schools. Districts B, C, and D each consist of an elementary, middle, and high school; while District A boasts a PreK-2 elementary building, a third through fifth grade intermediate school, a middle school, and a high school. In this, the districts we profiled appear similar to other districts using the TIGER observation model. A search of TIGER district websites suggests that participating districts range from one to seven schools in size. In 2017-18, the mean average daily membership among TIGER districts was 1,976 students (with TIGER districts' average daily membership ranging from 512 to 5,259) (TDOE, 2017-18 District Profile).

We wanted to understand whether patterns of responses to the 2018 Tennessee Educator Survey differed significantly between our districts or based on contextual factors, which might account for any variations in themes that may emerge from our interviews across districts.

[5]A district's TVAAS composite level ranges from level 1 (lowest performing) to level 5 (highest performing). It measures individual student growth using the Tennessee Value-Added Assessment System (TVAAS) (Tennessee Department of Education, TN Report Card, n.d.).

Participants

Teachers, school leaders, district administrators, and non-evaluative coaches comprised our primary qualitative sample. Because our research questions concerned staff's perception of teacher evaluation in their district context, we eliminated first-year teachers and staff new to the district from our sample. In each of the four districts, we requested the voluntary participation of one to two central office administrators, two school leaders (principals or assistant principals), two non-evaluative coaches, and four to five teachers from at least two of the district's schools (ideally an elementary and a secondary school). In many ways, we hoped to construct what Patton (2002) describes as a "typical case sample" by engaging the constellation of individuals involved in the TIGER observation protocol in order to "illustrate what is typical to those unfamiliar with the setting" (236). While our sample (described below in greater detail) largely did adhere to the guidelines we selected, and therefore includes elements of a typical case sample, it also reflects a heavy dependence on convenience sampling for efficiency's sake (Patton, 2002).

Rather than have members of the research team recruit individual participants in each district, we relied on our primary administrative contact to identify study participants on our behalf (and according to the guidelines we provided in terms of experience and position). We chose this approach both to maximize the efficiency of identifying participants and scheduling interviews and to maintain positive relationships with the district leaders we recruited into our study. Ultimately, we find that our sample includes a range of experiences and perspectives. Nevertheless, we recognize the limitations of our sampling approach (which we discuss more fully later), as it is possible that district leaders selected participants who they assumed would report favorably on the evaluation system, or excluded key informants due to scheduling or other constraints.

In each district, we engaged participants from two schools and the central office. In three districts, elementary (defined as grades PreK-5) and high school faculty, staff, and school leaders participated. In one district, middle and high school faculty, staff, and school leaders participated. For the purposes of reporting, we merge middle and high school into a "secondary" category and describe participants as either central office administrators, elementary faculty and school leaders, or secondary faculty and school leaders.

In total, seventeen teachers, five instructional coaches, nine school leaders (principals and assistant principals), and seven central office administrators participated in our study. On average teachers in our study boasted 13.97 years' experience and 11.41 years working in their respective districts. Teachers' overall experience ranged from two years to forty years, and experience in their district ranged from two years to thirty-nine years. Of the teachers in our study, 35.29 percent teach at the elementary level and 64.71 percent teach at the secondary level. School leaders, instructional coaches, and central office leaders who participated in our study claim an average tenure in their district of 13.95 years, with totals ranging from three years to 43 years.[6] (Note that tenure in the district does not necessarily equate to tenure in role).

Methods

We relied primarily on semi-structured interviews to gather our qualitative data about district context, patterns of TIGER implementation, and staff perceptions of the TIGER observation model. As Patton (2002) notes, interviews allow researchers to hear participants reflect on past experiences, offer interpretations that may illustrate broader concepts or themes, and assign meaning to events or experiences. Given the emphasis in our research questions on patterns, perceptions and context, interviews offer a trove of data to inform our understanding of individuals' experiences and identify patterns across individuals. To a lesser extent, we observed school and district environments, such as offices,

[6] One central office leader did not report years' experience in the district.

hallways, and classrooms. Our ability to observe school and district spaces was often limited by the constraints of our interview schedule and limited time in each district. We did not observe instruction, though we view this as only a minor limitation given that our interest lay in staff perceptions of evaluation, not the actual efficacy of instruction. We also examined documents and other artifacts related to evaluation. These included observation rubrics, classroom walkthrough guides, feedback forms, and TIGER observation model policy documents. (For a complete list of artifacts reviewed, see Appendix C). Our artifact analysis was limited to what documents district officials and participants were willing to share, and we did not collect consistent artifacts across each district (another limitation of our study).

Data Collection

We conducted a one-day site visit to each participating district during which we interviewed between nine and eleven district staff. To ensure a degree of consistency across interviews, we developed three interview protocols (see Appendix D): one for teachers, one geared toward school leaders (principals, assistant principals, and coaches), and one for central office administrators. We anchored our interview protocols in our conceptual framework and in themes that emerged from our extant literature review. As such, each protocol included questions related to school context and working conditions, administrative support and development, teacher collaboration, teacher evaluation, and observation and feedback. Connecting our interview protocols to our conceptual framework and extant literature review allows us to merge the insights from our review of prior research with what Patton (2002) describes as grounded theory, or theory that "emerges from the researcher's observations and interviews in the real world" (11). Ultimately, this results in a more nuanced understanding and interpretation of our research questions. We also asked questions about participants' background, such as tenure in the district. We did not ask participants to share their race or ethnicity, which may reflect a limitation of our study, given evidence that teachers of color receive lower evaluation ratings than white teachers (Campbell & Ronfeldt, 2018; Grissom, Bartanen, & Jones, 2018).

Most interviews took place one-on-one in private settings, such as a classroom or an office. A member of the research team conducted one interview in a semi-private setting, an office with a closed door but which contained a restroom that other individuals accessed at two points during the interview. We cannot discount the possibility that the semi-private nature of the interview space impacted the participant's candor. On four occasions, both members of the research team conducted an interview together. We must allow for the possibility that this created a perceived power imbalance that affected the interviewees' responses. Finally, in one instance, a central office leader joined the research team for lunch following his official interview with a research team member. This resulted in an additional, informal conversational interview with the district leader, which extended the research team's understanding of his perspective (Patton, 2002). Interviews typically lasted between thirty and sixty minutes, and we recorded interviews with participants' permission.

Between interviews, we observed classroom and office spaces and hallways, taking notice of the types of communications posted for staff, students, and parents, interactions among staff and between staff and students, and the general school or district environment. If interview participants shared artifacts, such as observation rubrics, with the team, we filed them away for subsequent review and analysis.

One team member also attended and observed two policy meetings relevant to our research questions. The first meeting, hosted by TERA in July 2018 convened TERA researchers, policy advocates, and members of Tennessee Department of Education to review Tennessee Educator Survey Data and other analyses of teacher evaluation in the state. State officials hoped to identify possible revisions to the TEAM evaluation model. In November 2018, the same team member attended a TIGER User Group meeting, a regular gathering of leaders from the state's TIGER districts for the purposes of examining and refining implementation of the observation model. At the November meeting, officials discussed best-practices in TIGER implementation from each district. These additional observations helped to clarify the department of education's interest in refining its policies in regards to evaluation and to inform our understanding of patterns of implementation across TIGER districts.

Data Analysis & Coding

As noted, the research team recorded all of our interviews. We did so using an application that simultaneously transcribed our conversations. In the first phase of analysis, we reviewed our interview recording and made any necessary corrections to the transcript. We then developed a thematic matrix for each individual interview. We organized each matrix around the concepts delineated in our conceptual framework and covered in our interview protocol (see Appendix E for an example). We then coded participants' responses according to these concepts, pulling out relevant quotations and, when applicable, citing additional evidence gleaned from observation or artifact analysis. After reviewing our codes, we generated themes related to each concept that emerged from the interview, thus beginning to induce a grounded theory from our fieldwork.

The research team then met to share findings from our individual interviews and analyses and to identify patterns and themes that emerged across our interviews and participating districts. This launched the second phase of coding and analysis, in which we developed a master matrix encompassing themes and findings from across all interviews. The research team organized this matrix according to themes we inductively generated from our interviews, observations, and artifact analysis and populated it with quotations and observations most relevant to each theme. From there, we concluded our qualitative coding and analysis by selecting the most salient quotes from each theme to include in the written description of our findings.

LIMITATIONS

We acknowledge a number of limitations across our study. While our quantitative analyses benefit from a large sample size and survey data generally recognized as valid, we found ourselves unable to conduct some relevant analyses due to limitations in the available data. For example, we could not control for student demographic data such as race, ethnicity, or English proficiency when comparing teacher survey responses — despite evidence suggesting that student demographics may play a role in teacher evaluations (Campbell & Ronfeldt, 2018). Additionally, our quantitative analyses do not take into consideration possible variations in responses based on teacher race/ethnicity, even though emerging research indicates that teachers of color receive lower ratings on evaluation systems than do white teachers (Grissom, Bartanen & Jones, 2018).

Our qualitative analyses include many more limitations. Our convenience sample relied on district leaders to identify teachers, coaches, and administrators to participate in qualitative interviews. Therefore, we cannot confidently claim that our sample of respondents is representative of the population of faculty and staff within each participating

district. Additionally, we visited schools selected by central office administrators, and it is possible that the themes we attribute to a district are more a reflection of a particular school context. These constitute threats to internal validity.

Our study also contains threats to external validity. First, we visited only four of Tennessee's eleven TIGER districts. While we believe these districts are largely representative of TIGER districts in terms of size and student population, we cannot say so for certain. It is possible that our findings reflect realities unique to the districts we visited as opposed to the TIGER observation model. Moreover, TIGER districts are uniformly small in size. The patterns of implementation that small districts can enact may not be relevant in larger school districts, raising questions about the applicability of our findings to other contexts. Furthermore, while our interview protocols adhered closely to our conceptual frameworks, they did not take into account demographic factors — such as respondent race/ethnicity or the income, race/ethnicity, disability status or language learner status of a teacher's students — that could influence respondent's answers to our questions (Campbell & Ronfeldt, 2018; Bartanen, Grissom, & Jones, 2018). Given that our primary research questions centered on patterns of implementation and teacher perceptions of observation and evaluation, we do not see the fact that we failed to collect information about the effectiveness of the teachers who participated in our study as a significant limitation. And of course, we must acknowledge our small sample size. For example, we interviewed only two percent of all teachers in participating districts.

Finally, a number of respondents in our study made comparisons between the TIGER observation model and the state's predominant TEAM evaluation. We report these comparisons; however, we cannot verify their accuracy as our study design did not include visits to any TEAM districts.

Despite these limitations, we sought to identify and triangulate rich sources of qualitative and quantitative data, and to approach our inquiries with discipline. We believe that the findings presented, discussion, and recommendations presented here provide fertile soil for further discussion. We hope that they support TERA and the TDOE in identifying avenues for further, more rigorous exploration and inquiry.

FINDINGS

Project Question 1: Is there a difference between teacher perceptions of evaluation in TEAM districts as compared to districts using an alternative observation model?

Finding 1A: Teachers report more favorable perceptions of evaluation as a tool for improving teaching in districts using alternative evaluation.

For each year (2015, 2016, 2017, and 2018) we computed z-scores to measure the difference in the proportion of educators from TEAM districts and from alternative evaluation districts who responded favorably to the two survey items of interest. We find that a greater proportion of teachers in alternative districts view the evaluation process as having led to improvements in their teaching ($p<.05$), leading us to reject the null hypothesis of no difference in proportion.

Conversely, between 2015 and 2017 we find no significant difference between the proportion of teachers suggesting that the feedback they receive is more focused on improving their teaching than making a judgment about their performance in TEAM districts and alternative evaluation districts, meaning we cannot reject the null hypothesis of no difference in proportion. In 2018, the proportion of teachers in TEAM districts who report that they receive feedback focused on improving instruction is greater than that of teachers in alternative evaluation districts ($p<.05$), leading us to reject the null hypothesis of no difference in proportions.

Table 4. Favorable responses to the survey item: "In general, the teacher evaluation process used in my school has led to improvements in my teaching."

TN Educator Survey Year	TEAM Districts	Alternative Districts	Total Population
2015	67%*	72%*	32,307
2016	71%*	75%*	27,835
2017	75%*	78%*	33,119
2018	73%*	75%*	35,250

*p < .05

Table 5. Favorable responses to the survey item: "During the current school year, the feedback that I received from my evaluator was focused more on helping me improve my teaching than making a judgment about my performance."

TN Educator Survey Year	TEAM Districts	Alternative Districts	Total Population
2015	44%	44%	31,117
2016	49%	47%	26,829
2017	54%	53%	32,981
2018	54%*	50%*	35,028

*p < .05

Finding 1B: Teachers report more favorable overall perceptions of evaluation and feedback in TIGER districts.

We conducted a z-test for the difference in proportions to compare favorable responses to our survey items of interest on the part of teachers in TEAM districts and teachers in TIGER districts in the years 2015, 2016, 2017, and 2018. Essentially, this duplicates our previous analyses but isolates the TIGER observation model.

We find that a greater proportion of teachers in TIGER districts view the evaluation process as having led to improvements in their teaching ($p < .05$), leading us to reject the null hypothesis of no difference in proportion. In 2016 and 2017, a greater proportion of teachers in TIGER districts report receiving feedback more focused on improvement than making judgments about their teaching ($p < .05$), leading us to reject the null hypothesis of no difference in proportion for these two years. These findings particularly are notable because this difference was not apparent when TIGER

districts were aggregated with other alternative evaluation districts. Taken together, these results suggest that teachers' perceptions of evaluation and feedback differed between TEAM districts and alternative evaluation districts, and between TEAM districts and TIGER districts.

Table 6. TEAM & TIGER Favorable responses to the survey item: "In general, the teacher evaluation process used in my school has led to improvements in my teaching."

TN Educator Survey Year	TEAM Districts	TIGER Districts	Total Population
2015	67%*	77%*	26,963
2016	71%*	82%*	24,570
2017	75%*	83%*	29,749
2018	73%*	81%*	30,584

*p < .05

Table 7. TEAM & TIGER Favorable responses to the survey item: "During the current school year, the feedback that I received from my evaluator was focused more on helping me improve my teaching than making a judgment about my performance."

TN Educator Survey Year	TEAM Districts	TIGER Districts	Total Population
2015	44%	47%	26,033
2016	49%*	53%*	23,711
2017	54%*	60%*	29,622
2018	54%	54%	30,390

*p < .05

Finding 1C: Within TIGER districts, district context is not significantly associated with teacher perceptions of evaluation and feedback — with the possible exception of teacher experience.

To better understand the potential influence of working in a TIGER district on teachers' responses to our survey items of interest, we conducted a series of ordinary least squares regressions that controlled for years of teaching experience, student demographics (as indicated by the proportion of a teacher's students eligible for Free and Reduced-Price Meals (FARMS), and teacher effectiveness (as measured by teacher value-added scores) using the 2017 and 2018 survey results (see Appendix F Table 10).

While these analyses do not offer a predictive model, they do indicate a systemic relationship between evaluation and feedback within TIGER districts. Our regression analyses showed no evidence that TIGER teacher responses vary after controlling for student demographics or teacher effectiveness. In contrast, experience may be related to teacher responses. For example, in 2018, for every one unit increase in teacher experience, teachers in TIGER districts are less likely to report that the feedback they receive is focused more on improvement than on making a judgment about their performance (ß = -.388; p<.0001). In 2017, for every one unit increase in teacher experience, teachers in TIGER districts are less likely to report that evaluation improves teachers (ß = -.011; p<.05) and less likely to perceive the feedback they receive is focused more on improvement than on making a judgment about their performance (ß = -.005; p<.05).

Therefore, in TIGER districts, contextual factors do not appear significantly associated with teacher perceptions of evaluation, with the possible of exception of teacher experience.

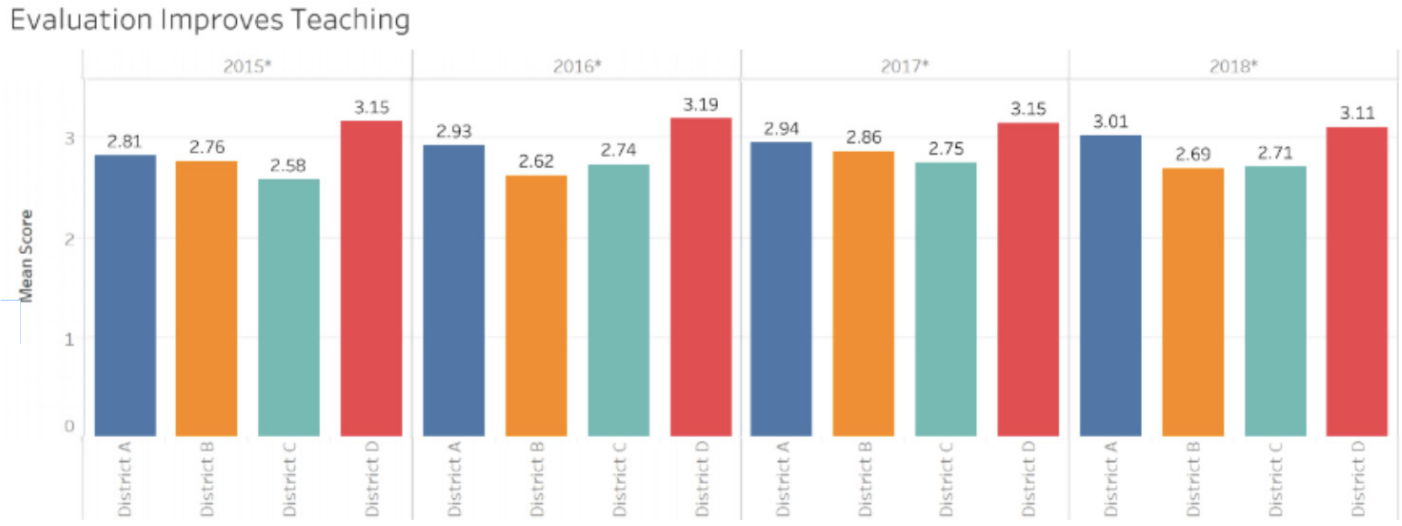
Target District Analyses

We wanted to understand whether our target districts varied in their patterns of responses to our survey questions of interest, as well as whether contextual factors appeared associated with teacher responses in these districts. These quantitative analyses of our target districts provide a point of comparison for the themes that emerge from our qualitative interviews.

Analysis of Variance

Accordingly, we conducted a one-way ANOVA to compare the effect of the target districts on the identified survey items between 2015 and 2018. In 2018, we found that teachers in District D responded more favorably to the question about whether evaluation improves teaching (p<.05; F (3,296) =5.53, p=.001). *Post hoc* comparisons using the Scheffe test indicate that the mean response from District D (M= 3.12, SD=.67) was significantly different from District B (M= 2.69, SD=.82) and District C (M= 2.71, SD=.81). A similar trend exists for survey years 2015-2017.

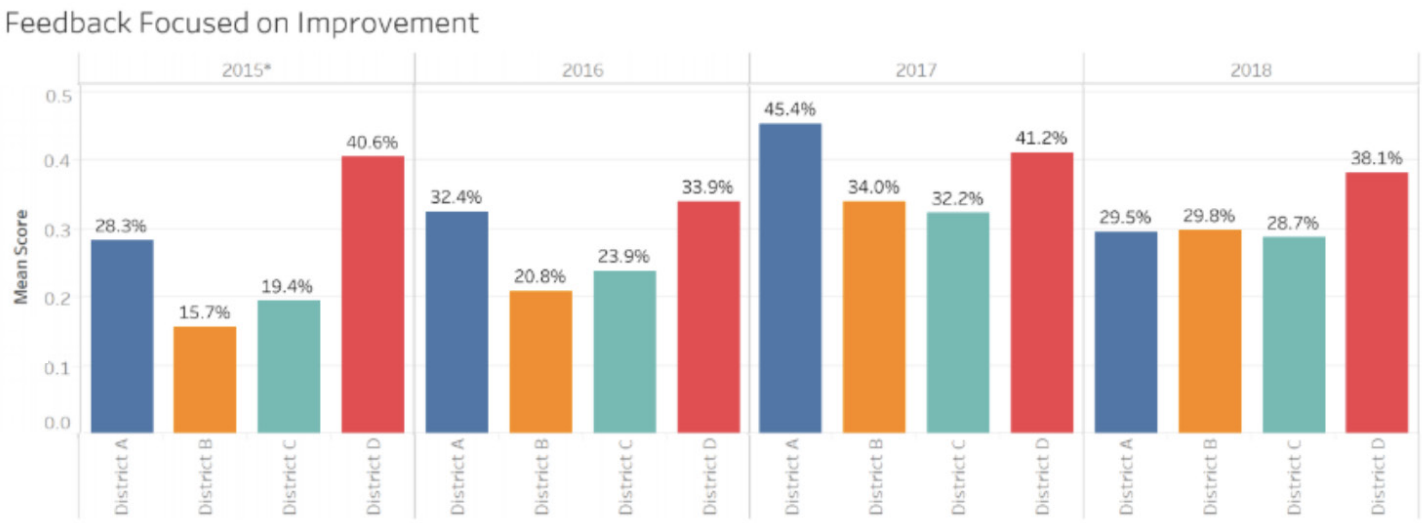
Table 8.



*Denotes statistical significance between schools in a given year (p<.01)

In 2015, teachers in District D were more likely to report that the feedback they receive is focused on improvement (p<.05, F (3,474) =7.96, p=0.001). *Post hoc* comparisons using the Scheffe test indicate that the mean response from District D (M= .41, SD=.49) was significantly different than District B (M= 2.69, SD=.82) and District C (M= .19, SD=.40).

Table 9.



*Denotes statistical significance between schools in a given year (p<.01)

Ordinary Least Squares Regressions

We conducted a series of ordinary least squares regressions using 2017 and 2018 survey results to understand:

- Whether favorable responses to our survey items of interest appear to vary in our target districts after controlling for contextual factors like teacher experience, student demographics (as measured by proportion of a teacher's students eligible for FARMS), and teacher effectiveness (as measured by TVAAS);
- Whether patterns of favorable responses to our survey items of interest appear to vary by schools within a target district.

With these analyses, we aimed to examine whether features of our sample (such as teacher experience or schools visited) may partially explain the themes that emerged from our interviews. Accordingly, we conducted a series of ordinary least squares regressions that used responses to our two survey questions of interest as the independent variable and controlled for:

- Target district;
- Teacher years of experience;
- Student demographics (as indicated by FARMS quartiles);
- Teacher Effectiveness (TVAAS level)

Because the data provided by the TN Educator Survey does not include information on student race/ethnicity or teacher retention rates, we were unable to control for these in our regression analyses.

We also conducted a separate regression that controlled for schools within a district to test the possibility that the particular schools we visited informed our perception of how stakeholders across a district perceive teacher evaluation.

Target District

Using District A as a reference category, we find some evidence of variability in teacher responses to our survey questions of interest between target districts. In 2018, teachers in District B (ß= -.327, p<.01) and District C (ß= -.304, p<.01) were less likely to report that evaluation improves their teaching. In 2017, teachers in District B (ß= -.114, p<.05) and District C (ß= -.131, p<.05) were less likely to report that feedback focused on improvement. However, these differences do not appear systematic across our limited two year analysis (e.g. we find no statistically significant difference in reports of whether evaluation improves teaching between districts in 2017 or whether feedback is focused on improvement in 2018). (See Appendix F Table 11).

Experience, Demographics, Effectiveness

We do not find compelling evidence that patterns of responses in our target districts appear related to teacher experience, demographics, or effectiveness.[7] In 2018, we saw no significant difference in responses across our target districts when controlling for teacher experience or for teacher TVAAS rating. In 2017, a one unit increase in teacher experience was associated with a decrease in favorable responses related to evaluation as a tool for improving teaching (ß=.081; p<.05) and a decrease in favorable responses relating to feedback focused on improvement (ß= -.050; p. <05). (See Appendix F Table 12).

Responses controlling for student demographics were largely similar in magnitude and direction. In 2018, teachers with 51-75 percent FARMS eligible students were more likely than those in districts with 0-25 percent FARMS eligible students to say report evaluation improved their teaching (ß = .625, p<.001), and in 2017 teachers with 76 –100 percent FARMS eligible students were less likely to report they received feedback focused on improvement (ß = -.058, p<.05). These results don't indicate systematic evidence of difference in our data based on student demographics. (See Appendix F Table 12).

Generally, these results suggest that contextual factors like teacher experience, TVAAS ratings, and student demographics are not substantively associated with patterns of responses in our target districts.

Schools

Within districts, response patterns vary by school. We conducted ordinary least squares regression analyses using 2017 and 2018 survey data that controlled for schools within a district. We found that in District B, teachers at the high school were significantly less likely to say that evaluation improves their teaching than teachers in the district's other schools (ß = -.516, p<.05). In the elementary school in District C, we found similar response patterns (ß = -.524; p<.05). In District A, teachers at the high school and elementary school were significantly more likely than teachers at the district's other schools to say that evaluation improved their teaching (ß = .501; p<.05; ß = .852; p<.01). And in District D, teachers at the middle school were significantly more likely than teachers in the district's other schools to say that the feedback they received is focused more on improving their teaching than making a judgment about their performance (ß = .518; p<.05). In Districts A and B in particular, these findings for specific schools resonated across our interviews in those buildings. (See Appendix F Table 13 & Table 14).

[7] Note, this finding differs slightly than what we found when we considered patterns of responses across all TIGER districts, in which we did see a possible association between teacher experience and perceptions of evaluation in both 2017 and 2018.

Project Question 2: How do district policies and school practices combine to influence patterns of implementation across districts using the TIGER observation model?

Finding 2A: TIGER's commitment to flexibility supports continued evolution of the observation model and contributes to varied implementation across contexts.

The TIGER model's menu of features provides district and school level leaders the flexibility to adapt observation and evaluation process to their specific context, which contributes to a commitment to model evolution and may explain variability in implementation across contexts.

Model Evolution

From the early stages of TIGER development, districts expressed a shared commitment to improving their schools by owning the educator evaluation process. One school director summarized the early conversations amongst AIMS directors. "Okay, if we're going to change teacher evaluation," the director recalled AIMS leaders thinking, "we want to have say, we don't want the state just to say this is this is the package deal, like it or not … We didn't want one size fits all."

This attitude is reflected in the considerable flexibilities built into the TIGER observation model and in continued collaboration among TIGER districts, whose leaders meet regularly as the TIGER User Group to examine and refine the TIGER approach. Member districts meet semi-annually to discuss variations in implementation across the eleven different TIGER districts and ensure that districts implement the non-negotiable core elements of the observation model with fidelity. As one TIGER district leader noted: "We have eleven different versions of implementation across the state. This meeting is a check-in to verify what is working well and learning from each other about variation in the process and local customization." This networked approach to improvement, which loosely resembles that advocated by Bryk, et al. (2015), has resulted in modifications to the overall TIGER observation model. For example, participating districts initially used a TIGER observation rubric (separate from the TEAM rubric) and an online reporting platform designed by the company Pearson. These proved difficult and costly to implement, leading TIGER schools to collectively agree to adopt the TEAM rubric and the state-sponsored TNCompass platform of reporting teacher performance. Leaders feel this change supported more streamlined implementation of TIGER without compromising the quality of the evaluation system. As one school principal remarked:

"I'm glad that TIGER switched over and is using the TEAM rubric because the TEAM rubric is good, it's solid. It is not that hard to make the switch from TEAM to TIGER. The most important aspect is that our beliefs about teachers' evaluation should also reflect how we think our students should learn. It's process-oriented."

Indeed, while AIMS leaders initially may have sought to develop an approach to evaluation distinct from the state's, TIGER users have evolved the observation model to function not so much as an alternative to the TEAM evaluation system as a set of observation and feedback processes and protocols that complement and enhance the TEAM system.[8] From the perspective of district leaders, TIGER functions as a robust observation and coaching model that simultaneously complies with state evaluation mandates. The flexibility of the model allows for variation in implementation and enables User Group members to monitor what approaches to implementation work best for whom and under what conditions.

Variation in Implementation

The flexible options within TIGER produce variation in implementation between and even within districts. The combination of coaching stages, state compliance requirements, non-negotiable core elements, flexible implementation options, and managerial discretion produce unique implementation models across observed districts and even across schools within a district. For example, some districts choose to use the TEAM rubric for formal observations (i.e. those that comply with state requirements) and content-specific protocols for the frequent walkthroughs and informal observations that characterize the TIGER approach. Some districts provide feedback via online platforms, while others prioritize in-person conversations between a teacher and observer, and some make greater use of peer observation than others.

While district leaders set broad priorities for schools in the district, we even find evidence of variation in implementation between schools in a district. Principals may choose to emphasize certain flexible options more than others or design specific management options, like scripts or forms, to support the observation and evaluation process. For example, in one elementary school, the principal and assistant principal elect to do all their observations together, with one scripting the lesson while another photographs student work and behaviors. The same team developed a unique platform using Google Drive to keep track of observations and feedback conferences (all of which they conduct in person with teachers). Another principal commented:

"TIGER does offer you some flexibility. I don't make them fill out all the forms that come in the packet because I think it's redundant, and I think it's a waste of their time to tell you the truth. But what we do make [teachers] fill out a growth plan, and they have to choose three reinforcements and three areas of refinement. And they have to elaborate on those and what they're going to do throughout the year, especially with the areas of refinement, to grow in those areas."

Thus, while district level leadership demonstrated a strong commitment to TIGER, building level leaders exerted significant influence in the variation of implementation within a district.

This flexibility seems to promote greater buy-in for the TIGER observation model. District leadership, building principals, and even teachers comment on how they have used this flexibility to promote professional learning. For

[8] For this reason, and given that TIGER districts use the TEAM rubric and state reporting system, we are careful to refer to the TIGER observation model as opposed to the TIGER evaluation system. It could very well be argued that TIGER districts merely complement the TEAM evaluation system with a varied observation model.

example, one district leader describes the freedom to prioritize practices to shape school culture:

> *"The hoops are fewer in TIGER. I don't have to do the one announced, unannounced full lesson plan. My time has to add up to an hour. Y'all, that's awesome. There's a lot of freedom because what are you trying to do is positively impact instruction, right? Well, there is no one way to do that… using that flexibility how it best serves you and your schools need, TIGER gives you the freedom to do that. To adapt [observation and evaluation] … it's a culture, it's not a piece of paper."*

In other instances we observed that variation across contexts contributed to different perceptions of evaluation. For example, in a district that relied primarily on online feedback as opposed to in-person conversations, we heard more mixed reviews of evaluation and saw less evidence that teachers make robust use of formative feedback. As one elementary school teacher in the district said, "Yeah, I mean, I read [the online feedback]. And if there is a question or a problem, you know, I do want to take care of that. I don't want to just leave that out there. But it's just kind of evidence of the moment and yeah, move on. I don't dwell on it."

This raises questions about whether all variations that districts adopt support improved implementation of the TIGER observation approach.

Finding 2B: TIGER's holistic approach to evaluation builds buy-in from teachers and administrators.

As we discuss more fully in additional findings, the TIGER observation model does not score individual observations. This is considered a non-negotiable to which all participating districts agree to adhere. Instead, teachers participating in the TIGER observation model receive a summative evaluation at the end of the year that reflects the entirety of evaluator observations, classroom walkthroughs, perceived engagement with professional development and the school community, and other factors. To some, this might imply a high degree of subjectivity in teacher summative ratings. As one secondary school principal noted, "When we get to the end of the year, and it's time for us to sit down and start assigning scores, you know, we've got all that [data from across the year]. Some of it is documented, some of it's not. Some of it's just what we've seen, what we have in our head, and that does persuade us when we start assigning scores to a teacher." Indeed, this subjectivity could contribute to biases in evaluator ratings, and had we collected data on respondent race or ethnicity, it is possible that a different picture of the TIGER observation model might have emerged.

Nevertheless, we found that teachers and administrators almost universally appreciated this approach to assigning ratings, which we dubbed a holistic evaluation approach. Principals we interviewed felt that the holistic approach allowed them to give more nuanced, accurate ratings of teachers than simply averaging scores across classroom observations. For example, one elementary school principal observed that the combination of formal observations and consistent, informal walkthroughs provided "...lots of snapshots to put together to have this big album at the end of the year instead of just a couple pieces." Teachers appreciated feeling like their summative annual rating would reflect the entirety of their professional practice, and that one misstep during a formal observation would not have undue impact on their ratings. One elementary teacher claimed she breathes easier because "the evaluation process is not 'I'm getting evaluated today. And it is the end all be all.'" Another secondary school teacher, with prior experience in a TEAM district, reflected on the differences in her evaluation experiences:

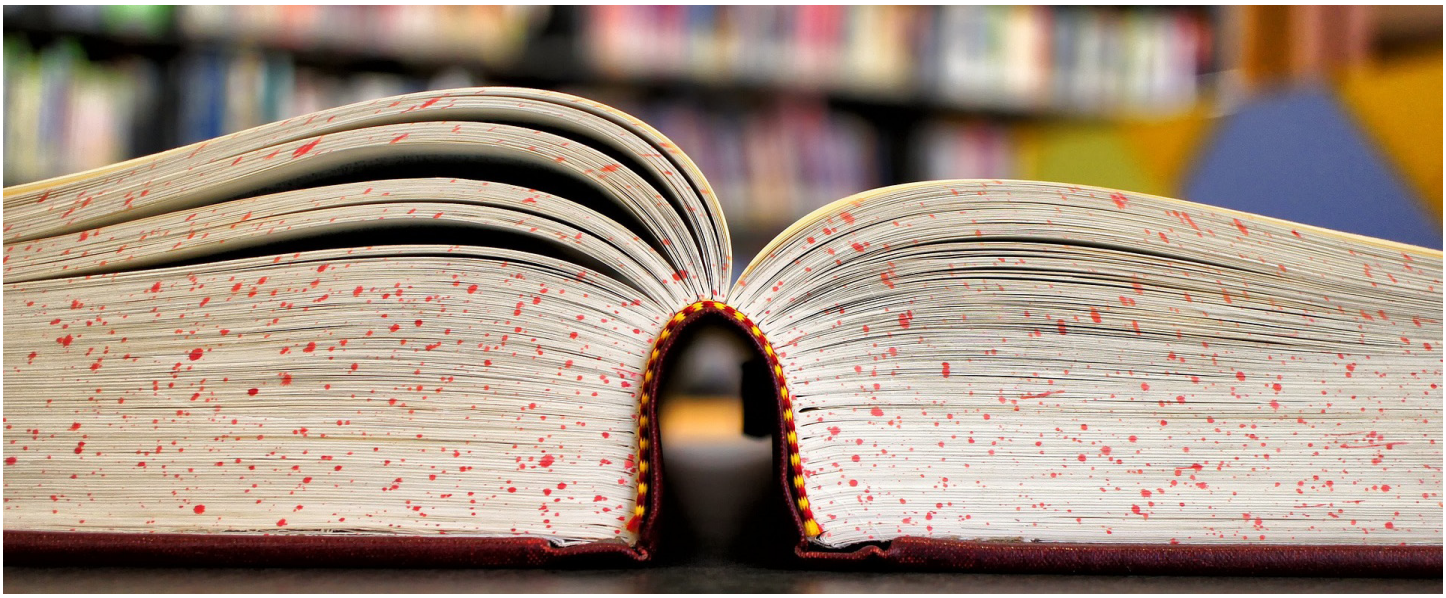
> *"We did not have TIGER in my previous district. And I was observed one time in the fall and one time in the spring. And it was it almost felt like you better hope that those are your best days teaching because that's the score that you're going to get. Whereas here, it's a lot of summative data collecting where it's like, okay, we're going to come in at different times of the day with different classes. And we're going to see a whole picture instead of just a snapshot."*

Teachers and administrators collectively appeared more bought in to the TIGER model thanks to its emphasis on multiple factors and holistic scoring, suggesting that the approach improves perceptions of evaluations.

Nevertheless, we did encounter some limits to the subjectivity and holistic nature of the rating. For example, more than one principal and central office leader across our districts reflected on the importance of teacher TVAAS scores. These leaders shared that they regularly analyze state TVAAS data and compare it to their ratings of teachers, and that differences between TVAAS ratings and summative evaluation ratings raise concerns. According to one secondary principal, "If you have a teacher that consistently, year after year, level one with her TVAAS, there's a problem, and you have to pay attention to that. So I think it's important." In another district, a secondary principal baldly stated: "If [a teacher's] value added falls below a three, then we're gonna put them in [Stage I]." And in a third district, the Director of Schools described sharing TVAAS ratings alongside principal evaluation ratings across the district to generate a conversation about discrepancies. This suggests that even as administrators value the relative autonomy and subjectivity of the holistic evaluation approach, they still look to more objective measures to corroborate their ratings.

Another potential, though less clear cut, limitation to the holistic evaluation approach relates to TIGER's principle of negotiation. In theory, TIGER allows for "negotiated evaluation" in which the evaluator and teacher together determine the teacher's summative rating. In practice, we found that evaluators — most often principals — typically drove this conversation. Some teachers described engaging in a degree of negotiation around their final scores. For example, one fourth-year teacher described arriving at her summative rating in the following, collaborative way: "I assess myself. She'll assess me. We'll sit down and talk, and where we are good we go on and move through, and where we differ — whether I give high and she gives low, or she gives high and I give low — we talk about it. I'm able to justify. She's able to justify, and then we find the meeting ground."

However, we much more commonly heard from both teachers and administrators that administrators determine the final score with minimal input from teachers. In the words of one secondary principal, "Very rarely do I negotiate [a final score]. [My AP] and I assign the scores, and if the teachers question it we have that conversation, 'well, let's look back at the rubric, this is why we think you fell in this area.' And normally the teachers don't argue with them." Interestingly, teachers appeared unperturbed by this non-negotiated approach. In fact, none of the teachers we interviewed raised concerns or expressed frustration if they did not have a chance to negotiate a final score. We might conjecture from this that the generally holistic nature of the summative ratings built greater trust in the administrator's final assessment.



Finding 2C: While central office leaders evinced a commitment to maintaining a firewall between evaluation and coaching, in practice, that firewall is sometimes porous.

Among TIGER's key features is an apparent commitment to separating evaluation and coaching. In theory, Stage I teachers or any teacher requiring additional support should receive targeted feedback and guidance from a non-evaluative coach who does not share observations about the teacher in question with the teacher's evaluator. In the districts we visited, central office leaders in particular vocally endorsed this principle. One central office leader who also coaches teachers stated plainly: "I have no role in teacher evaluations because coaches are not evaluators." An elementary school principal in another district reinforced the firewall, saying: "We don't want our coaches or teacher partners to be perceived as evaluators. And so they're strictly feedback and support."

Nevertheless, the TIGER observation model calls for clear lines of communication between evaluators (most often principals) and coaches for the purposes of identifying a teacher's coaching needs. "So the coach and the principal and the teacher sit together," one evaluation coordinator explained, "and they come up with the growth plan." In the process of executing that growth plan, coaches must then communicate to principals if and when a teacher is ready for another, formal observation. Theoretically, this is the extent of the communication. The coach understands from the principal what the teacher's perceived needs are. The coach and teacher confidentially address those needs, and the coach tells the principal when the teacher has improved sufficiently to warrant a next observation.

In reality, teachers perceive that coaches and principals communicate about teachers' progress. In fact, teachers appear to struggle to differentiate in some instances between a coach and an evaluator. A secondary teacher described frequent conversations between "the evaluators" (in this case, she was referring to a central office employee who serves as a coach and her principal, or formal evaluator):

> *"Because I know, they get together, the different evaluators get together and they talk about [their observations] as well. Are you seeing what I'm seeing? And that helps a lot too, because you're not getting fifteen different things that you have to work on. They're focusing on the main thing that you have to work on."*

It is important to note that in this instance, the teacher valued the apparent collaboration between the coach and the evaluator, because she felt it provided her with more consistent feedback and support.

The line between formative walkthroughs and formal observations also seemed fuzzy. A veteran secondary teacher with twenty-two years' experience reflected: "As far as the evaluations go — they're in and out constantly. So you really never know, are they just popping in because they want to sit here are they popping in to do a formal evaluation. But they are, like I've said before, really good about giving you the feedback." Again, the teacher appears unconcerned by the confusion because she appreciates the quality of the feedback she receives regardless of the visit's purpose.

It is possible that teachers we spoke with seemed not to see the same bright line between coaching and evaluation as central office leaders and school administrators because most of the districts we visited emphasized the principal's role as an instructional leader and encouraged principals to visit classrooms frequently and provide feedback to teachers. According to one director of schools: "We've had some workshops with our principals and assistant principals on what good feedback looks like, what it sounds like, how to recognize what the standards really are." A principal in another district described working diligently to provide content-based, constructive feedback to all teachers in the building.

Moreover, in three of the four districts we visited both teachers and principals reported that principals frequently visit classrooms for the purposes of providing feedback. "We're in our classrooms much more than any district around us,

I can tell you that," one secondary principal noted. It is not impossible therefore, that a teacher sees her principal and evaluator walking through her classroom more frequently than she does her assigned coach. As principals continue to step into instructional leadership roles, they may naturally find themselves engaged in coaching teachers. Indeed, one district evaluator admitted, "When I'm an evaluator, I still try to keep that coaching hat on, too." This may contribute to confusion about the separate roles of principal and coach, but that confusion does not appear to influence teachers' perceptions of the TIGER observation model or the feedback they receive.

In one district (perhaps not incidentally the same district that employs online feedback protocols), this ambivalence on the part of teachers to the separation between coaching and evaluating did not manifest as clearly. In this district, especially at the secondary level, teachers reported fewer walkthroughs on the part of their principals. Therefore, administrative observations felt more uniformly evaluative in nature, and even somewhat threatening. One veteran teacher described evaluation as "another obstacle to navigate," noting, "I don't know that I've ever had somebody just come in and go 'Hey, just seeing what you're doing." The relative infrequency of administrative observations led some (though not all) teachers in this district to believe that most teachers "are not themselves" when administrators enter their classrooms. An elementary school teacher in the district described evaluation as "still something that I kind of fear and get nervous and worried about."

While limited in scope, these findings imply that the degree to which teachers see school leaders and evaluators as instructional leaders with deep knowledge of their teaching may influence how much weight they place on the supposed firewall between coaching and evaluation.

Finding 2D: One of the TIGER observation model's strengths lies in its emphasis on teacher growth, as opposed to compliance or accountability.

The clear majority of interviewees positively described TIGER as a growth-oriented approach to evaluation. As one elementary principal observed, "It's a growth process and into growth model. And that is what I love about the TIGER model. It is not a gotcha, or, you know … we're going to take all of this and we're just going to average it up. It is a growth model. So it's not where you begin, it's where you end." Both early career and veteran teachers reflected this in their comments. As one -year veteran teacher noted, "I think the TIGER model in general has done a good job of just trying to find ways to improve without breaking down teachers." While some teachers and administrators acknowledged a secondary aim of evaluation as identifying and dismissing ineffective teachers, none identified that as the primary purpose of teacher evaluation in their contexts.

By and large, respondents attributed TIGER's focus on growth to two features of the observation model — its emphasis on feedback and its holistic approach to assigning summative ratings. We touch more deeply on both of these elements in findings 2E and 2B. Here, we note that teachers appear to credit the frequency and the quality of the feedback they receive through TIGER as a major component in their own development. Far from reducing the rigor of evaluation, TIGER's emphasis on growth and formative feedback may push teachers to accelerate the pace of their own professional improvement. According to one secondary teacher:

> *"You are going to be pushed to your limits teaching here ... but the reward is greater here than it is in other districts. You're pushed hard, but you get a bigger reward. We have lots and lots of observations here, walkthroughs. They [administrators] give feedback, they give critiques. We get critiques every week. We have professional development every month. They are constantly pushing you to make better questions, to incorporate different activities, to push those limits of your students."*

Other interviewees noted that deemphasizing scores throughout the year helps reorient teachers and administrators away from accountability and toward growth. "Every time you give a score," one secondary principal posited, "it

becomes about the score and not about the feedback." Teachers also appeared more comfortable with the evaluation system given its emphasis on growth and feedback, as opposed to frequent scoring.

One teacher who previously worked in a TEAM district made the following comparison:

> *"TEAM to me is a little more intimidating because you are, it feels more like an A, B, C, D, F is what TEAM feels like. TIGER, when I came over here to TIGER. TIGER does not feel as stressed. It's much more like a conversation."*

Another secondary teacher with prior experience in a TEAM district described scored observations as a "cart and pony show" focused on compliance and performance, whereas in TIGER "the feedback is much more natural, it's much more helpful, and it's much more on target than a cart and pony show."

The observation model's emphasis on growth appeared to contribute to cultures of continuous improvement within participating districts. A number of interviewees reflected sentiments such as "we all have areas where we can improve in" (central office leader), which seemed to defuse some of the pressure of a high-stakes evaluation system. Indeed, we saw observational evidence of this growth mindset in some of the buildings we visited. For example, in one secondary school teachers post their area of refinement outside their classroom doors, alongside an invitation to their colleagues to observe their practice and provide feedback. All of the districts we visited described embedding frequent opportunities for teachers to collaborate around instruction, through structures like coach-facilitated, weekly professional learning communities, daily collaborative planning, or monthly early release days. With few exceptions, commitments to professional development and public practice permeated the districts we visited and contributed to cultures of improvement.

Finding 2E: "Just enough, Just in Time" feedback promotes improvement in TIGER districts.

The TIGER observation model goes above and beyond state policy requirements in its emphasis on frequent observation and feedback cycles. District leaders are focused on professional learning and for instructional improvement. As one director of schools stated:

> *"We have moved from a single event to a continuous and ongoing coaching professional development model. Our professional development has to be just enough just in time based on where our needs are. That can change from week to week."*

A commitment to deprivatizing teaching practice in most of the schools we visited contributed to the development of this feedback culture.

<u>Public Practice & Observation</u>

We saw schools work to deprivatize teaching by prioritizing open and transparent practice, as evidenced by frequent peer, coach, and administrator observation. When asked about the support and development of teachers, one instructional coach responded:

> *"My job to grow teachers, that's my job. And in order to grow teachers, you need to be visible and with them, they need to feel comfortable when I come in their classroom because I'm in there all the time."*

Although the TIGER process manual calls for at least two walkthroughs per semester for Stage II/III teachers, in practice we found that most schools far exceed this recommendation. In November, one teacher reported having

received at least a dozen classroom walkthroughs since the beginning of the school year. One school director aims to have every teacher receive twenty to thirty walkthroughs each year. Another secondary teacher described peer observation as "something we try to do anyway, just try to see what's going on, and see if there's some other way to teach another technique that works." And, as previously noted, one school went so far as to have teachers post their areas of refinement on their classroom doors.

This culture of public practice connects to an intentional effort we observed in many districts to support teachers in improving their instruction. According to one instructional coach: "We do so many walkthroughs… the biggest support here is just the amount of feedback teachers get throughout the year." And teachers appreciate the increased specificity of feedback that districts are pushing evaluators to provide. "There is more feedback and it is more specific," reflected an elementary school teacher. "Before they were coming in and I didn't know what they were looking for, now I have more of an idea of what they are looking for and what I am trying to improve."

While the quote above illustrates sentiments communicated in many of our interviews, we did encounter a circumstance in which teachers describe a more tenuous relationship between observation and growth. In the district we have previously described as somewhat of an outlier, principals seemed to focus more on complying with guidelines (e.g. the number of minutes spent in classrooms) as opposed to the value of observations and walkthroughs for formative development. Teachers picked up on this nuance:

> *"Ideally, [TIGER] is a growth model with lots of feedback," said one teacher, "This is mostly the case for Stage I teachers… More and more it has become an objective report about what happened. If you need timely feedback to make a change, you [the evaluator] have to be there [with feedback] tomorrow. That is the difference between calling something a growth model and making it a growth model."*

In this district, teachers could articulate the value of TIGER "in theory," yet they remained apprehensive about the support they received and seemed more reluctant than teachers in other participating districts to deprivatize their teaching.

Where teachers did receive frequent feedback, it appeared most impactful when closely coupled with coaching and support.

Coaching, Feedback, and Support

Most of the TIGER schools we visited prioritized providing rapid feedback to teachers and integrating it with targeted coaching and professional development opportunities for teachers. Central office administrators and school leaders stressed that this feedback be closely linked to a teacher's practice, either using the TEAM rubric or more content-specific tools like Student Achievement Partners Instructional Practice Guides (IPG) (which multiple stakeholders mentioned they prefer to TEAM's more broad guidance). Some schools and districts also oriented feedback around a coherent, shared focus area — such as student engagement — that served as an anchor for the year's professional development. Regardless of the observation tool or focus area, most TIGER districts we visited emphasized the need to marry feedback with action. "Coaching is where we make our hay," one director of schools evocatively noted. An instructional coach elaborated on this, observing:

> *"In order for this process to be implemented, to be a growth model, the person who's evaluating has to give you something to do. I feel like that's a sure way of getting buy-in for any sort of evaluation system is ensuring that people who are using the tool can provide feedback that will help you grow."*

This shift toward actionable feedback and related supports appeared particularly evident to teachers with prior experience in TEAM districts. "In the previous district, when I was just with TEAM," recalled one teacher, "I didn't have the coaching model or that coaching aspect to it. It was just: 'This is what you need to work on. Go find it your own. Go find it yourself.'" In contrast, a teacher noted, the feedback she received in her TIGER district was complemented by the ongoing presence of a mentor or coach "to help us find different strategies to align to our professional development."

Ultimately, we observed the ongoing commitment to observation, feedback, and coaching to be a hallmark of the TIGER observation model in the districts we visited. Even in our outlier district, where implementation appeared spotty, leaders cogently described and seemed philosophically committed to these elements of the TIGER approach.

Project Question 3: How does school and district context influence perceptions of teacher evaluation in districts using the TIGER observation model?

Finding 3A: Contextual factors like district size and quality of relationships may influence perceptions of the TIGER observation model and raise questions about the feasibility of scaling it.

Context appeared to influence perceptions of the TIGER observation model among our interviewees. In particular, we noticed that the relatively small size of the TIGER districts we visited contributed to a sense of tight relationships within and across districts, which respondents frequently cited as a reason for the observation model's success in their context.

No district in our study houses more than four schools. A number of respondents credited the size of their districts as a key factor in successful implementation of the observation model. "You know, you're talking about collaboration," one central office leader said, reflecting on TIGER, "A small district has the ability to do that."

As noted previously, most TIGER districts are fairly small ones, and a number of our interviewees questioned whether TIGER could successfully be implemented in larger school districts. One director of schools observed that all of Tennessee's TIGER districts combined serve fewer students than the Knox County School district. The director expressed reservations about wide scale implementation of TIGER absent some sort of oversight and support function on the part of the state. A central office leader in another district shared:

> *"I don't know how comfortable I would be with that kind of flexibility [that TIGER provides] if I was in charge of 49 schools … Sometimes when you have that many people, it can really get convoluted. My interpretation. Your interpretation. We can talk that out here, you know, what does that look like … However, it is with that flexibility that me and these administrators … we know our teachers pretty well, and we can work this [TIGER] depending on where you are and what you're doing and maybe be able to meet needs a little better, because it's not so structured."*

The small size of TIGER districts and the sense that therefore "everyone knows everyone" may contribute to the strong relational culture that appeared to exist in many of the districts we visited. For example, one elementary teacher said the following about her building administrators: "Because I have a personal relationship with them, I know that when they walk into the room, they are going to see the best in me. And they're also going to see things in me that they want truly to help me to reflect on to get better." Other teachers shared that they trust their colleagues to collaborate with them and support their improvement. A secondary teacher said the following about her school's professional culture:

"Our relationships in this building may be unique. I know in some places it's hard to find a place where everyone comes together, and they share their ideas. But we do, for the most part. I can shoot [another teacher] an email and ask her about something or just catch her in the hallway … and then anything that she had worked on, she sends me. So they're very willing to share … It's really good in this building the way teachers share and collaborate."

Some of the administrators and central office leaders we interviewed described undertaking intentional work to build such collaborative, trusting cultures. Sometimes this took the form of strategically retaining long-time district staff. In other instances, leaders described intentionally leveraging (or encouraging) turnover to create cultures in which everyone shares the same vision for success. "In the five years I've been here," shared one principal, "there's been a huge turnover. So most of the teachers are teachers I've hired. And that goes back to what I said before — the culture of the building — I wanted to establish a strong culture." Regardless of how this culture gets built, if respondents are correct in inferring that the small size of TIGER districts may facilitate greater relationship building, trust, and collaboration among faculty and staff, then it raises questions about the feasibility of scaling the TIGER observation model to other, larger districts across the state of Tennessee.

Finding 3B: Districts in which stakeholders spoke most favorably of the TIGER observation model invested in differentiated and ongoing supports for teachers, administrators, and coaches.

Central office leaders often spoke of the TIGER observation model less in the context of high-stakes teacher evaluation and more as an opportunity to inject much needed routines around feedback, support, and instructional improvement in their districts. “I was really interested in building a culture here,” reflected one such leader, “where teachers saw the need for better professional development to kind of improve their craft.” Accordingly, most of the districts we visited, respondents described a number of differentiated support structures that augmented the TIGER observation model. As previously noted, one district appeared an outlier in this regard, in that teachers and principals described fewer such supports and also expressed more mixed opinions of teacher evaluation. In this district, some teachers, though not all, made comments like, “We talk about changing the way we teach, but in the meantime, we don’t get the tools.” While we cannot discount these more critical opinions, most of the respondents we interviewed spoke highly of the professional supports provided by their districts.

Coaching represents perhaps the most obvious form of support associated with the TIGER observation model. Within TIGER, all Stage I teachers receive additional coaching supports, and most districts followed the flexible option of making coaching available for any teacher at any stage. According to one Stage II teacher, “We have math and reading language arts coaches that are supportive and give us different avenues and tools to help us differentiate and work on our lessons.” Additionally, the TIGER districts we visited invested in varied support structures beyond coaching.

In the case of supports for teachers, interviewees described regular engagement with professional learning communities (PLCs) and frequent district professional development days, such as monthly “Power Fridays.” In most districts, these appeared focused on the successful implementation of standards and curriculum to support instruction. Ensuring that these collaborative structures remained focused on topics relevant to instruction emerged as a priority in many interviews. Some districts tapped instructional coaches to lead PLCs. Another district held teachers accountable for developing clear meeting agendas and turning in a work product following each collaborative session. “The lead teacher has to provide us with an agenda,” shared a coach in that district, “so that we know exactly who's going to be there, what the topic is, what the work product's going to be from there ... we have to have a work product. You know, just going in and everyone having a chat session is not a work product.” Regardless of the approach the district took to building accountability for professional learning, we heard and saw evidence in most buildings that teachers engage in frequent, focused coaching and professional development.

In addition to supporting teachers, districts invested in supporting principals, coaches, and administrators in leveraging the TIGER observation model to build instructional leadership skills. Whereas districts often led teacher development efforts in-house, they seemed to rely more heavily on external partners to support school leaders’ and instructional coaches’ development. In two districts, respondents cited TDOE’s Read to be Ready program as a key support in developing coaching skills. “Read to be Ready,” said one coach, “I mean, that's my biggest support right now, as a coach, because we look at the phases of coaching, you know, and then I have a regional coach that comes at least once a month, sometimes twice. We have regional meetings, usually once a month ... so as a coach, that is my biggest support system right now.” Another coach described participating in multiple coaching networks through partners like Lipscomb University. And three of the four districts we visited described some connection with the nonprofit Instruction Partners, which focuses on ensuring that students access grade-level content and rigorous instruction (Instruction Partners, School and District Partnerships, n.d.). Coaches in particular credit these forms of professional development with improving their own practice. “We looked at three or four [videos of me coaching]” one shared of an external professional development session, “I saw right off where I was doing a good job and where I was a failure. It doesn't bother me to be a failure if I can fix it.”

The prevalence of these support structures raised questions about whether the TIGER observation model in and of itself is sufficient to drive instructional improvement, or whether its success depends on nesting it within a constellation of coaching, professional development, and other collaborative supports for teachers and leaders. Given this possibility, some leaders expressed a desire for the state to provide some degree of support for districts implementing TIGER. According to one director of schools, "I really want to beef up the training, I would love there at the state office to be somebody who's in charge of TIGER. That would be my number one wish."

Finding 3C: Building leaders play a key role in implementing TIGER and establishing cultures of professional growth in their buildings.

In the districts we examined, district leaders often empowered building principals to adapt TIGER to meet the needs of their faculty. In buildings where teachers seem most engaged in TIGER as a professional growth tool, they view members of the school administration as their building's primary instructional leaders. As one elementary school teacher put it, "We look to our administrators. They are very good, very knowledgeable in that [instructional leadership] respect." Principals who engaged in frequent walkthroughs, provided targeted feedback to teachers, and intentionally aligned professional learning with school instructional goals appeared most successful in catalyzing a growth culture within their buildings.

In three schools (two secondary and one elementary) across two districts, we encountered teachers who didn't identify TIGER participation as a critical factor in their growth (even if they did note that TIGER was intended to support growth). In these cases, we were left with the impression that principals focused more on meeting TIGER's compliance mandates than on leveraging its flexible options to support ongoing teacher development. In these buildings teachers more commonly pointed to peers as their primary source of instructional leadership. "We have a great [grade level] team," said a teacher in one of these buildings, "there are six of us. We work very well together. And most all of us are what I would consider very, very hardworking."

Even in these instances, district leaders recognized the significance of the principal role, acknowledging a need to better support them in serving as instructional leaders. According to the district's Director of Schools:

> *"At the district level, from our point of view, we've had to provide a lot for our principals, for our evaluators. We've had some workshops with our principals and assistant principals on what good feedback looks like, what it sounds like, how to recognize what the standards really are."*

These two contrasting examples both point to the role the school leader plays in setting the tone for TIGER implementation in their building.

DISCUSSION

The findings across our quantitative and qualitative analyses of districts implementing the TIGER observation model in Tennessee largely align with existing and emergent research literature related to teacher evaluation, instructional improvement, and the contextual factors that enable or inhibit the scale of effective practice. Thematically, we orient our discussion around the following observations, all of which overlap with elements of our conceptual framework:

1. The TIGER observation model appears to have successfully oriented itself toward what Papay & Richard (2018) describe as the second goal of teacher evaluation reform - "improving instructional practice through feedback and support" (2).

2. Successful implementation of the TIGER observation model, and likely any teacher evaluation approach, requires substantial investment in building the instructional capacity of teachers, administrators, and coaches.
3. Contextual factors related to district size and relational trust likely influence the degree to which the TIGER observation model can be effectively scaled.

Observation 1: The TIGER observation model appears to have successfully oriented itself toward what Papay & Richard (2018) describe as the second goal of teacher evaluation reform — "improving instructional practice through feedback and support" (2).

In their examination of teacher evaluation as a tool for instructional improvement, Papay & Richard (2018) write: "[A]lthough nearly all states explicitly list professional learning as the goal of evaluation reform, implementation efforts … have struggled to create systems at scale that enable principals and teachers to translate evaluation feedback into improved instruction" (Papay & Richard, 2018). Districts using the TIGER observation model may buck this trend. Our analyses suggest that teachers in TIGER districts are more likely than other teachers in Tennessee to report that the evaluation process improves their teaching.[9] In part, this may reflect the ways in which TIGER district leaders have deemphasized the accountability elements of teacher evaluation in favor of opportunities for teacher development. Indeed, a number of stakeholders noted that they see the evaluation component of the TIGER observation model as an exercise in "compliance" for the state, but that they leveraged the evaluation requirements to institute processes and practices that emphasize coaching, feedback, and development. According to one director of schools, "[Evaluation] does have a human capital piece to it. It needs to have a human capital piece to it. But it also has to have a professional development piece to it, and it's two different buckets. And here's what's great about the TIGER model is that what we're trying to do is separate the two conversations." Our qualitative findings suggest a number of factors that may contribute to this successful decoupling of the compliance and development aspects of evaluation within TIGER. These include the TIGER observation model's emphasis on holistic scoring, the deprivatization of practice in participating TIGER districts, and an apparently growing emphasis on content-based feedback for teachers.

First, TIGER's emphasis on holistic scoring may mitigate some of the negative perceptions of teacher evaluation as a high-stakes accountability tool. Many stakeholders we interviewed expressed positive feelings about the TIGER observation model as a tool for growth, not a 'gotcha' accountability system. As one high school teacher put it: "You do not get a score on the rubric until the end of the year. If you have a bad observation in August, there is still a lot of time to recover." This emphasis on growth and selecting a holistic, summative rating at the end of the year (as opposed to averaging the scores of each classroom observation) may serve to lower the stakes of classroom observations throughout the year. Grissom & Loeb (2017) report that principals are more likely to provide differentiated performance ratings in low-stakes evaluation settings, suggesting that teachers in TIGER districts may receive more targeted feedback about their performance throughout the year because the observation model delays the high-stakes rating pressure until the end of the school year.

Additionally, the observation model's emphasis on the use of multiple data points — including formal observations, informal classroom walkthroughs, and other evidence of instruction or engagement in the school community — appears to increase the amount of times evaluators and coaches observe teachers' instruction. A number of respondents spoke to the ways in which the TIGER observation model reinforces the deprivatization of teaching practice. According to one sixth year teacher: "So far, this is our fourteenth or fifteenth week of school, and I've

[9] Admittedly, teacher perception survey data suggest that early career teachers perceive the most benefit from the supports surrounding the TIGER observation model. This aligns with empirical research suggesting that, although teachers can and do improve throughout their careers, improvements are often largest in the first 10 years (Kraft, Papay, & Chi, 2018; Papay & Laski, 2018; Sun, Mutcheson & Kim, 2015). It also may speak to TIGER's particular utility as an observation model for early career teachers.

probably had fifteen observations. At least one a week, sometimes two a week." The normalization of observation on the part of administrators, instructional coaches, and even other teachers in most of the TIGER districts we visited likely contributes to cultures of professional improvement.

Bryk, et al. (2010) argue that an orientation toward continuous improvement and professional community are two key indicators of a school's overall professional capacity. At the core of a strong professional community, they write, are "work arrangements for faculty that (1) make teachers' classroom work public for colleagues and external consultants; (2) institute processes of critical dialogue about classroom practice … and (3) sustain collaboration among teachers that focuses on the school's instructional guidance system" (Bryk, et al., 2010, 56). In their longitudinal analysis of school outcomes in Chicago, Bryk, et al. (2010) find that professional capacity functions as an essential support that contributes to sustained improvement in student achievement and attendance. Writing more recently, Papay & Richard (2018) link observation (i.e. public practice) to the development of teachers' professional capacity, reporting that teachers improve more rapidly in buildings in which observations occur frequently, are conducted by multiple people, and which provide differentiated feedback to teachers.

Finally, our qualitative interviews revealed that, to varying extents, participating districts are leveraging the TIGER observation model and its encouragement of frequent walkthroughs to emphasize content-specific feedback and coaching. As one principal explained: "You know, it's developed over time. It used to be a very simplistic [walkthrough rubric] that applied to every subject. Now we use, we have a[n] ELA walkthrough, a math walkthrough." The principal later emphasized that he sees the "heavy lifting" of feedback and support coming in the walkthroughs, which he uses as a "major piece" of data when assigning a teacher's holistic score "because it's much more rigorous than the TEAM model [sic] in terms of [being] much more specific to a subject." In other districts, teachers spoke of having math and ELA specific coaches available to offer content-based tools, resources, and feedback.

The TIGER districts we visited may be unique in their emphasis on using classroom walkthroughs and coaching to provide content-based feedback. Nationally, the trend seems to favor more generic observation approaches. Hill & Grossman (2013) decry one-size-fits-all teacher observation instruments since "research generally indicates that

subject-specific professional development and coaching are more effective in improving instruction." An American Institutes for Research (AIR) study reinforces this view, finding that generic observation rubrics provide teachers with little practicable feedback on their instruction (Welch, et al., 2016). Of the forty-five teacher evaluation observation instruments the AIR team reviewed, only eight were subject-specific (Welch, et al., 2016). This is despite the fact that teachers appear to convey more legitimacy on subject-specific evaluation systems. A U.S. Department of Education commissioned study of teacher evaluation systems finds that teachers respond more readily to evaluation feedback when they trust the observer's content expertise (Anderson, Butler, Palmiter, & Arcaira, 2016). Our interviews confirmed this perception among teachers. In one participating district, which did not appear to emphasize content-based feedback as much as others in the study, a secondary teacher with eight years' experience commented: "I mean if you have somebody observing you that's never taught [your subject], I don't know how they can really help you change your teaching or give you nuggets of feedback when they've not taught the subject before."

To be clear, the TIGER observation model itself does not emphasize content-based feedback. Indeed, it uses the same, generic observation rubric as most Tennessee districts. Instead, we find district leaders leveraging the general flexibility TIGER provides to embed content-specific walkthroughs and coaching into district routines, procedures, and supports. This, in turn, may enhance the utility and credibility of the observation model among stakeholders. This emerging focus on content-based feedback, combined with TIGER's commitment to holistic scoring, coaching, and ongoing observation (i.e. more public practice), create a general sense among stakeholders in participating TIGER districts that the observation model emphasizes growth over accountability.

Observation 2: Successful implementation of the TIGER observation model, and likely any teacher evaluation approach, requires substantial investment in building the instructional capacity of teachers, administrators, and coaches.

For teacher evaluation approaches to facilitate improvement, they must be coupled with targeted efforts to build the capacity of teachers in response to evaluation findings. The TIGER observation model's explicit emphasis on coaching speaks to this imperative. Beyond coaching, interviewees in most of the TIGER districts we visited described a number of structures, routines, and procedures aimed at improving teachers' instructional skills. These included professional learning communities, regular staff development days (e.g. a monthly early release), and peer observation and coaching. In the districts where these collaborative structures appeared most embedded, teachers described feeling well-supported in their craft. "I've got my principal, I've got my supervisor, I've got my coach, I've got my co-workers," enthused one elementary teacher, "I mean there's just a plethora of people that I could go to help me."

Research suggests that efforts to encourage professional collaboration among faculty will contribute to school improvement. In a finding echoed by Murphy & Torre (2014), Bryk, et al. (2010) describe quality professional development as a "key instrument for change," especially when it occurs in the context of a generally supportive professional environment and a coherent curriculum (125). Other scholars reinforce these conclusions, noting that teachers improve at greater rates when they work in buildings that, among other things, prioritize quality professional collaboration and development (Papay & Kraft, 2017; Ronfeldt, Owens Farmer, McQueen, & Grissom, 2015). Indeed, evidence from Tennessee's own Instructional Partnership Initiative indicates that teachers can powerfully motivate improvement among their peers (Papay, et al., 2015). However, in most districts it appeared that administrators and instructional coaches often guided these collaborative efforts among teachers, contributing in some cases to a sense of fatigue. "We have a lot on us," said one secondary principal, "in terms of what we have to do as administrators." The potential for administrator burnout, combined with powerful evidence about the potential impact of high-quality professional collaboration among colleagues, raises questions about the extent to which TIGER districts make use of the Stage III teacher designation.

In the districts we visited, it appeared that teachers rarely receive Stage III designation, which the TIGER procedural manual describes as a recommended, but optional element of the TIGER observation model (AIMS/TIGER 2017). Stage III intends to provide highly effective teachers with opportunities for peer leadership and coaching, often married with additional pay. The TIGER manual stipulates that in districts using Stage III, teachers must apply to receive the designation. Administrators we spoke with referenced budgetary constraints and limited teacher interest in the application process as key barriers to the more widespread adoption of the Stage III designation. Nevertheless, we wondered if more strategic use of the Stage III designations could enable districts to further embed norms around professional collaboration without overextending administrators and coaches.

Finally, a contrasting example speaks to the potential importance of investing in teachers' professional capacity by promoting greater collaboration among faculty. In our outlier district teachers struggled to identify opportunities to learn with and from peers. "We've talked about common planning time before," bemoaned one veteran teacher, "and I think it just becomes kind of a logistics nightmare for the people that do the scheduling." Another experienced teacher in the district lamented limited opportunities to observe other teachers. It may not be surprising that in a district where teachers report relatively sparse professional development and collaboration activities, teachers evince a more skeptical view of teacher evaluation.

TIGER places a burden on administrators, especially building leaders like principals, assistant principals, and instructional coaches, to mitigate such negative reactions by creating the conditions for high-quality collaboration and feedback. As one director of schools observed: "There are basically three people at the central office who have anything to do with instruction … everything else is done at the building level." Given this emphasis on school building leadership, TIGER districts must invest heavily in these individuals' development and capacity to lead instructional teams and provide high-quality feedback to teachers.

Research consistently affirms the critical role that school leaders play in promoting and sustaining school improvement. Bryk and colleagues (2010) speak to "the centrality of principal leadership initiating and sustaining the organizational changes necessary to improve student learning. Quite simply," they write, "school improvement is highly unlikely to occur in its absence" (205). Leithwood, Seashore Louis, Anderson, & Wahlstrom (2004) famously report that school leader quality (behind teacher quality) represents the second most important in-school factor in student learning. Leaders support student learning in part through their ability to develop faculty instructional competency, providing clear direction, and creating organizational structures that promote effective instruction (Leithwood, et al., 2004). Indeed, teachers in our study look to building leaders to provide this kind of guidance. "I have the support of, number one, my principal," reflected one secondary school teacher with four years' experience. "He's great to sit down and talk to and gives me advice."

In order to play this pivotal instructional leadership role, principals, assistant principals, and instructional coaches must receive sufficient supports themselves. Central office leaders in the districts we visited appeared to recognize this imperative. For example, in our outlier district, where we observed less evidence of principal instructional leadership than in others, the central office had recently hired a staffer with deep expertise in coaching and leadership development, and the district's Director of Schools cited principal development as a key area of need. In other districts, we saw that external providers, such as the state department of education, universities, and nonprofits, play an important role in supporting principals and coaches. Central office leaders viewed these partnerships as important for developing staff and a critical resource for small districts with limited central office capacity. "You know, it's just me and one other gal," explained a district supervisor of teaching and learning, "and she does special populations. So we really look to [our external partners]" to provide high quality resources and professional development.

Despite the reliance on external partners to provide support to principals and coaches, central office leaders universally advanced a commitment to instructional leadership and saw the TIGER observation model's emphasis

on feedback as key to doing so. As one director of schools succinctly said, "I believe in the power of coaching and feedback … so I've tried to build my coaches up so that they can coach teachers." Research suggests that these districts' emphases on providing support for administrators as instructional leaders likely contributes to strengthened building leadership and better implementation of teacher evaluation. Although she cautions against contracting with external providers to develop building leaders, Honig (2010) posits that when executive-level leaders elevate the importance of instructional leadership, it is more likely to permeate district practice. Grissom, et al. (2017) noted that when principals received strong central office support around talent management functions like teacher evaluation, they felt better equipped to undertake these key instructional leadership roles than principals in districts without such robust supports. These findings echo those of Steinberg & Sartain (2015), who saw that teacher evaluation contributed more greatly to teacher improvement when evaluators received significant capacity-building support from their central office. In brief, the majority of TIGER districts we visited put in place robust support structures for school leaders, instructional coaches, and teachers to build their instructional capacity, likely contributing to perceptions of the TIGER observation model as a tool for improving teacher practice.

Observation 3: Contextual factors related to district size and relational trust likely influence the degree to which the TIGER observation model can be effectively scaled.

TIGER districts are modest in size, and stakeholders across our interviews consistently referred to the tight network of relationships that naturally occur in smaller districts. "Now, we're very small here," went one characteristic comment from a central office leader. "We all know each other pretty well. So a lot of times in most buildings, just because I walk in [to a classroom] and sit down, it doesn't mean 'Oh, they've sent somebody from the central office over here.' [Teachers] know me well enough to know I want to be in those rooms, okay. I'm a support system for them." Teachers in three of the four districts we visited echoed these comments, describing largely collegial, trusting relationships with other teachers and with administrators. These comments suggested to us that the TIGER districts we visited are, for the most part, characterized by a high degree of relational trust, and evidence suggests that such trust often forms the basis for successful school improvement (Bryk, et al., 2010; Murphy & Torre, 2014; Bryk & Schneider, 2002). According to Bryk & Schneider (2002), "a broad base of trust across a school community lubricates much of a school's day-to-day functioning and is a critical resource as local leaders embark on ambitious improvement plans" (5). Additional literature indicates that this trust may be more likely to emerge in small school environments, which Coleman (1988) might refer to as a closed social structure. This research posits that because small schools are characterized by dense social interactions that facilitate the development of social capital, or norms of reciprocity, familiarity, and trust among community members (Torre & Murphy, 2016; Coleman, 1988).

Our interviews suggest that teachers' generally positive perceptions of the TIGER observation model stem in part from this reservoir of trust in their administrators and colleagues. As one twenty-two-year veteran teacher with experience in multiple districts put it, "It's very, very enjoyable to teach here. The faculty, there's not the cliques that go on with the faculty. The administration is, and this is probably the biggest thing, the administration is extremely supportive here." These findings raise questions about whether the TIGER observation model, with its emphasis on frequent walkthroughs, observation, and authentic feedback, could be successful in districts without such reserves of trust.

Elmore (2004) raises a similar caution. He observes that the success of systems-level initiatives depends entirely on their ability to penetrate practice "in the smallest unit — the classroom, the school" (5). Indeed, Elmore (2004) hypothesizes that successfully scaling good educational practice may require both that teachers operate within a system of relationships that provide training and support and that systems tighten the traditionally loose coupling between administrative structures and the core technology of classroom practice. In small systems characterized by robust social capital and trust, system leaders likely wield a greater understanding of and influence over the work of the school and classroom. This, in turn, facilitates the successful implementation of the TIGER observation model, which, with its focus on observation, coaching, and feedback, is highly relational. Whether this can be feasibly accomplished in larger, less socially networked districts remains to be seen. As our findings report, even TIGER system leaders remain skeptical of its utility in larger contexts. However, we note that approximately 30 percent of Tennessee school districts have average daily membership counts at or below the mean TIGER district average daily membership – and 70 percent of districts in the state have average daily membership counts at or below the that of the largest TIGER district – suggesting that even as size may be a consideration in a district's decision to adopt TIGER, a number of districts within Tennessee fall well within the range of TIGER districts' average daily membership (TDOE, District Profile 2017-18). Ultimately, our observations indicate that if state or system leaders sought to expand the use of the TIGER observation model, or even elements of it, they would need to pay considerable attention to incentivizing a greater emphasis on growth over compliance, to developing mechanisms to build the capacity of teachers and administrators, and to the social context in which they deploy the observation model.

RECOMMENDATIONS

From a policy perspective, Tennessee benefits from a consistent teacher evaluation system in TEAM that allows for comparability across the majority of the state's districts. As TERA supports the state in considering whether and how to refine TEAM, the TIGER observation model provides an example of an approach that, by and large, successfully supports a growth mindset. While TIGER's successful implementation seems to rely on deep reservoirs of relational trust between teachers and administrators, some features of the observation model may more readily be translated across contexts. To that end, we recommend that TERA and TDOE consider:

Holistic scoring of teacher evaluations. TEAM guidelines currently call for evaluators to determine a teacher's summative score by averaging the teacher's scores across each observation (Hunter 2018). In the TIGER districts we visited, we heard that teachers and administrators valued the opportunity to assign a summative score based on the teacher's performance across the year and taking into account multiple forms of evidence. While it could be argued that taking an average score across observations accomplishes this, stakeholders we interviewed felt that not scoring each observation allowed teachers and evaluators to focus on growth. Moreover, the permission TIGER gives to consider formal observations in addition to informal walkthroughs and professional development in assigning summative scores led interviewees to believe the observation model provides a more accurate reflection of teacher skills, building buy-in for the system. Importantly, the focus on holistic scoring does not appear to negatively impact overall student growth in our target districts. Only one district in our study had a 2017-18 TVAAS level below a three, and that maybe an aberration (perhaps due to a leadership transition in a district school), as the district's composite TVAAS level ranged from three to five between 2013 and 2017 (Tennessee Department of Education, TN Report

Card, n.d.). Therefore, we posit that adopting a commitment to holistic scoring represents a relatively painless, but potentially impactful, refinement to the TEAM evaluation system.

Replacing the generic TEAM rubric with content-specific rubrics. While all TIGER districts use the TEAM general educator rubric for teacher evaluations, we heard from stakeholders in some districts that for informal walkthroughs they prefer using content-specific Instructional Practice Guides from external partners such as Instruction Partners. In particular, administrators with whom we spoke felt that content-specific rubrics enabled them to provide more specific feedback to teachers. Moreover, content-specific rubrics could mitigate some of the concerns teachers in our study (and across the extant literature) raise regarding their evaluators' content expertise (Hill & Grossman, 2013; Kraft & Gilmour, 2016; Donaldson, 2013). Our review of research bolstered our findings, as it suggested that teachers are more likely to improve in response to content-based, as opposed to generic, feedback (Hill & Grossman, 2013). While TDOE already provides observation guidance specific to special populations of teachers or subjects (such as special educators, early childhood educators, or K-3 literacy), these could be strengthened by revising the actual general educator observation rubric to reflect the content being taught (TDOE, 2018, Evaluator Handbook). At minimum, we suggest that Tennessee consider providing TEAM rubrics specific to English Language Arts/social studies and math/science. Ideally, these rubrics also would take into consideration grade bands such as elementary, intermediate, and secondary.

Supporting regular, regional convenings of district leaders to examine and improve teacher evaluation implementation. One of the most compelling features of the TIGER model is the User Group network comprised of district and school leaders. The User Group focuses on monitoring TIGER implementation and sharing practices related to the flexible management options within the TIGER model. The User Group functions like an improvement community dedicated to understanding TIGER implementation in participating districts, examining adaptations by district context, and articulating improvements to the overall model (Bryk et al., 2015). We recommend TDOE consider supporting regional networks of TEAM district leaders who meet to examine and improve the implementation of teacher evaluation across districts. These meetings could support more coherent implementation of the model, build expertise and shared accountability among participants, and unearth opportunities for TDOE to allow for greater adaptation and flexibility related to supporting instructional improvement. In turn, these may maintain or build buy-in for evaluation among district leaders.

Building mechanisms for non-evaluative coaching into the evaluation system. TEAM provides clear guidance for conducting post-observation conferences in ways that promote teacher reflection and identify clear areas for development (TDOE, 2018, Evaluator Handbook). Less clear is whether TEAM explicitly encourages or requires any sort of coaching for educators. By contrast, the TIGER observation model requires that Stage I teachers receive non-evaluative coaching, and our observations and interviews suggested that many Stage II teachers also take advantage of opportunities to work with coaches. This investment in coaching reflects research suggesting that coaching impacts teachers' instructional improvement (Hill & Grossman, 2013; Kraft, Blazar, & Hogan, in press). Moreover, it reinforces TIGER's commitment to leveraging teacher evaluation for growth and improvement by providing teachers with protected space to develop their skills. We recommend that TEAM incorporate more widespread access to protected coaching, at least for early career or less effective teachers.

We provide this recommendation with some caveats. First, providing widespread access to coaching requires significant human capital investments on the part of districts or the state. Second, most of the TIGER districts we visited appear to benefit from high concentrations of relational trust, making teachers seemingly less concerned about maintaining a clear separation between coaching and evaluation. In districts that do not benefit from such tightly knit relationships, administrators may need to maintain a clearer firewall between evaluators and coaches in order to build teacher buy-in to the coaching process. Finally, as we shall discuss, coaches would need to have strong capacity to diagnose and support instruction. This may require districts, or the state, ramping up supports for coaches across Tennessee.

Incentivizing more public practice and prioritizing timely, actionable feedback within and beyond formal observations. The TIGER observation model deemphasizes formal observations for accountability purposes in favor of encouraging frequent walkthroughs and other, informal opportunities to observe and provide feedback on teacher practice. The prioritization of public practice and informal observation contributed to cultures of growth within many of the buildings we visited. Currently, TEAM allows informal walkthroughs to augment teacher evaluation only for level five educators (TDOE, 2018, Evaluation Handbook). The system might be strengthened and more oriented toward growth if state leaders encouraged the use of frequent walkthroughs for teachers at all levels of effectiveness. Because making time for walkthroughs and other forms of informal observation may prove challenging for principals, the state may want to incentivize such behavior by letting walkthroughs for all teachers count toward required observation minutes or replacing one formal observation with three walkthroughs across the year (Kraft & Gilmour, 2016; Donaldson, 2013). That being said, our findings indicate that public practice most impacts perceptions of evaluation when it leads to more timely, actionable feedback for teachers

Stakeholder interviews suggested that when teachers receive frequent feedback from administrators and coaches (and even peers), they feel well-supported and confident in their growth as educators. The districts we visited uniformly prioritized building school leaders' knowledge and skills around diagnosing instruction and providing feedback. Those districts taking action on this front typically looked to external partners, rather than in-house expertise, to build their leaders' and coaches' capacity. This suggests a possible role for TDOE. Emerging research from TERA and elsewhere is exploring the impact of feedback on instructional improvement (Erin O'Hara, TERA, personal communication; Springer, Gegenheimer, Hunter, & Koedel, forthcoming). As those findings come to light, we encourage TDOE to consider creating ongoing professional development aimed at supporting school leaders and instructional coaches in providing content-specific, actionable feedback to teachers. This could be incorporated into existing supports, like the Tennessee Academy for School Leaders, or reflect new offerings geared jointly to administrators and coaches (TDOE, n.d, Tennessee Academy for School Leaders). A number of stakeholders we interviewed spoke highly of TDOE professional development offerings, such as Read to be Ready, suggesting that TDOE may be better positioned than some other state education agencies to provide capacity building supports to educators.

We recognize that our final two recommendations depend heavily on competent school leaders and coaches as well as effective cultures within schools and districts, making them more difficult to enact at scale. Nevertheless, our field research and our review of the extant literature suggest that these serve as critical enabling conditions for the successful implementation of any educational improvement (Bryk, et al., 2010; Murphy & Torre, 2014; Kraft & Papay, 2014; Papay & Kraft, 2017). For that reason, we argue that the state must pursue opportunities to build capacity among school and district leaders as part of any attempt to further strengthen the TEAM evaluation system.

CONCLUSION

State and federal policies continue to prioritize teacher evaluation as a lever to improve teacher effectiveness and bolster student achievement. Since 2012, Tennessee has served as a national model for the successful implementation of teacher evaluation. Recent data suggests that not only does the state lead the nation in student growth, but that teachers generally see the value in evaluation as a tool for improvement (Reardon, 2018; Papay & Laski, 2018). For this, the state deserves commendation.

At the same time, our field research indicates that teachers and administrators largely perceive the TEAM evaluation system as an accountability mechanism that attempts to graft an improvement orientation onto itself. By contrast, the stakeholders we interviewed generally perceive the TIGER observation model as a growth-oriented tool for

teacher improvement that also provides a means of complying with teacher evaluation mandates. These perceptions stem from TIGER's emphases on holistic scoring, observations, feedback and coaching, and flexible paths to implementation.

Our findings suggest that as TDOE looks to maintain the state's status as a national leader in teacher evaluation and to improve the statewide implementation of TEAM, it may need only look in its own backyard. In particular, if TDOE, with TERA's support, hopes to evolve TEAM to focus first on improvement, and second on accountability, TIGER provides a powerful model of such an approach.

APPENDIX A

TEAM General Educator Rubric

General Educator Rubric: Environment

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Expectations	• Teacher sets high and demanding academic expectations for every student. • Teacher encourages students to learn from mistakes. • Teacher creates learning opportunities where all students can experience success. • Students take initiative and follow through with their own work. • Teacher optimizes instructional time, teaches more material, and demands better performance from every student.	• Teacher sets high and demanding academic expectations for every student. • Teacher encourages students to learn from mistakes. • Teacher creates learning opportunities where most students can experience success. • Students complete their work according to teacher expectations.	• Teacher expectations are not sufficiently high for every student. • Teacher creates an environment where mistakes and failure are not viewed as learning experiences. • Students demonstrate little or no pride in the quality of their work.
Managing Student Behavior	• Students are consistently well behaved and on task. • Teacher and students establish clear rules for learning and behavior. • The teacher overlooks inconsequential behavior. • The teacher deals with students who have caused disruptions rather than the entire class. • The teacher attends to disruptions quickly and firmly.	• Students are mostly well behaved and on task, some minor learning disruptions may occur. • Teacher establishes rules for learning and behavior. • The teacher uses some techniques, such as social approval, contingent activities, and consequences, to maintain appropriate student behavior. • The teacher overlooks some inconsequential behavior, but at other times, stops the lesson to address it. • The teacher deals with students who have caused disruptions, yet sometimes he or she addresses the entire class.	• Students are not well behaved and are often off task. • Teacher establishes few rules for learning and behavior. • The teacher uses few techniques to maintain appropriate student behavior. • The teacher cannot distinguish between inconsequential behavior and inappropriate behavior. • Disruptions frequently interrupt instruction.
Environment	The classroom: • welcomes all members and guests, • is organized and understandable to all students, • supplies, equipment, and resources are all easily and readily accessible, • displays student work that frequently changes, and • is arranged to promote individual and group learning.	The classroom: • welcomes most members and guests, • is organized and understandable to most students, • supplies, equipment, and resources are accessible, • displays student work, and • is arranged to promote individual and group learning.	The classroom: • is somewhat cold and uninviting, • is not well organized and understandable to students, • supplies, equipment, and resources are difficult to access, • does not display student work, and • is not arranged to promote group learning.
Respectful Culture	• Teacher-student interactions demonstrate caring and respect for one another. • Students exhibit caring and respect for one another. • Positive relationships and interdependence characterize the classroom.	• Teacher-student interactions are generally friendly, but may reflect occasional inconsistencies, favoritism, or disregard for students' cultures. • Students exhibit respect for the teacher and are generally polite to each other. • Teacher is sometimes receptive to the interests and opinions of students.	• Teacher-student interactions are sometimes authoritarian, negative, or inappropriate. • Students exhibit disrespect for the teacher. • Student interaction is characterized by conflict, sarcasm, or put-downs. • Teacher is not receptive to interests and opinions of students.

APPENDIX A

TEAM General Educator Rubric

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Standards and Objectives ☐	• All learning objectives are clearly and explicitly communicated, connected to the state standard(s), and referenced throughout lesson. • Sub-objectives are aligned and logically sequenced to the lesson's major objective. • Learning objectives are: (a) consistently connected to what students have previously learned, (b) known from life experiences, and (c) integrated with other disciplines. • Expectations for student performance are clear, demanding, and high. • There is evidence that most students demonstrate mastery of the daily objective that supports significant progress towards mastery of the standard(s).	• Most learning objectives are communicated, connected to the state standard(s), and referenced throughout lesson. • Sub-objectives are mostly aligned to the lesson's major objective. • Learning objectives are connected to what students have previously learned. • Expectations for student performance are clear. • There is evidence that most students demonstrate mastery of the daily objective that supports significant progress towards mastery of the standard(s).	• Few learning objectives are communicated, connected to the state standard(s), and referenced throughout lesson. • Sub-objectives are inconsistently aligned to the lesson's major objective. • Learning objectives are rarely connected to what students have previously learned. • Expectations for student performance are vague. • There is evidence that few students demonstrate mastery of the daily objective that supports significant progress towards mastery of the standard(s).
Motivating Students ☐	• The teacher consistently organizes the content so that it is personally meaningful and relevant to students. • The teacher consistently develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher regularly reinforces and rewards effort.	• The teacher sometimes organizes the content so that it is personally meaningful and relevant to students. • The teacher sometimes develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher sometimes reinforces and rewards effort.	• The teacher rarely organizes the content so that it is personally meaningful and relevant to students. • The teacher rarely develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher rarely reinforces and rewards effort.
Presenting Instructional Content ☐	Presentation of content always includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • effective modeling of thinking process by the teacher and/or students guided by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; and • no irrelevant, confusing, or non-essential information.	Presentation of content most of the time includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • modeling by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; and • no irrelevant, confusing, or non-essential information.	Presentation of content rarely includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • modeling by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; and • relevant, coherent, or essential information.

APPENDIX A

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Lesson Structure and Pacing ☐	• The lesson starts promptly. • The lesson's structure is coherent, with a beginning, middle, and end. • The lesson includes time for reflection. • Pacing is brisk and provides many opportunities for individual students who progress at different learning rates. • Routines for distributing materials are seamless. • No instructional time is lost during transitions.	• The lesson starts promptly. • The lesson's structure is coherent, with a beginning, middle, and end. • Pacing is appropriate and sometimes provides opportunities for students who progress at different learning rates. • Routines for distributing materials are efficient. • Little instructional time is lost during transitions.	• The lesson does not start promptly. • The lesson has a structure, but it may be missing closure or introductory elements. • Pacing is appropriate for less than half of the students and rarely provides opportunities for students who progress at different learning rates. • Routines for distributing materials are inefficient. • Considerable time is lost during transitions.
Activities and Materials ☐	• Activities and materials include all of the following: o support the lesson objectives, o are challenging, o sustain students' attention, o elicit a variety of thinking, o provide time for reflection, o are relevant to students' lives, o provide opportunities for student-to-student interaction, o induce student curiosity and suspense, o provide students with choices, o incorporate multimedia and technology, and o incorporate resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.). • In addition, sometimes activities are game-like, involve simulations, require creating products, and demand self-direction and self-monitoring. • The preponderance of activities demand complex thinking and analysis. • Texts and tasks are appropriately complex.	• Activities and materials include most of the following: o support the lesson objectives, o are challenging, o sustain students' attention, o elicit a variety of thinking; o provide time for reflection, o are relevant to students' lives, o provide opportunities for student-to-student interaction, o induce student curiosity and suspense; o provide students with choices, o incorporate multimedia and technology, and o incorporate resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.). • Texts and tasks are appropriately complex.	• Activities and materials include few of the following: o support the lesson objectives, o are challenging, o sustain students' attention, o elicit a variety of thinking, o provide time for reflection, o are relevant to students' lives, o provide opportunities for student to student interaction, o induce student curiosity and suspense, o provide students with choices, o incorporate multimedia and technology, and o incorporate resources beyond the school curriculum texts (e.g., teacher made materials, manipulatives, resources from museums, etc.).

General Educator Rubric: Planning

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Instructional Plans ☐	Instructional plans include: • measurable and explicit goals aligned to state content standards; • activities, materials, and assessments that: o are aligned to state standards, o are sequenced from basic to complex, o build on prior student knowledge, are relevant to students' lives, and integrate other disciplines, and o provide appropriate time for student work, student reflection, and lesson unit and closure; • evidence that plan is appropriate for the age, knowledge, and interests of all learners; and • evidence that the plan provides regular opportunities to accommodate individual student needs.	Instructional plans include: • goals aligned to state content standards, • activities, materials, and assessments that: o are aligned to state standards, o are sequenced from basic to complex, o build on prior student knowledge, and o provide appropriate time for student work, and lesson and unit closure; • evidence that plan is appropriate for the age, knowledge, and interests of most learners; and • evidence that the plan provides some opportunities to accommodate individual student needs.	Instructional plans include: • few goals aligned to state content standards, • activities, materials, and assessments that: o are rarely aligned to state standards, o are rarely logically sequenced, o rarely build on prior student knowledge, and o inconsistently provide time for student work, and lesson and unit closure; and • little evidence that the plan provides some opportunities to accommodate individual student needs.
Student Work ☐	Assignments require students to: • organize, interpret, analyze, synthesize, and evaluate information rather than reproduce it, • draw conclusions, make generalizations, and produce arguments that are supported through extended writing, and • connect what they are learning to experiences, observations, feelings, or situations significant in their daily lives both inside and outside of school.	Assignments require students to: • interpret information rather than reproduce it, • draw conclusions and support them through writing, and • connect what they are learning to prior learning and some life experiences.	Assignments require students to: • mostly reproduce information, • rarely draw conclusions and support them through writing, and • rarely connect what they are learning to prior learning or life experiences.
Assessment ☐	Assessment plans: • are aligned with state content standards; • have clear measurement criteria; • measure student performance in more than three ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); • require extended written tasks; • are portfolio based with clear illustrations of student progress toward state content standards; and • include descriptions of how assessment results will be used to inform future instruction.	Assessment plans: • are aligned with state content standards; • have measurement criteria; • measure student performance in more than two ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); • require written tasks; and • include performance checks throughout the school year.	Assessment plans: • are rarely aligned with state content standards; • have ambiguous measurement criteria; • measure student performance in less than two ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); and • include performance checks, although the purpose of these checks is not clear.

APPENDIX A

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Grouping Students ☐	• The instructional grouping arrangements (either whole-class, small groups, pairs, individual; heterogeneous or homogenous ability) consistently maximize student understanding and learning efficiency. • All students in groups know their roles, responsibilities, and group work expectations. • All students participating in groups are held accountable for group work and individual work. • Instructional group composition is varied (e.g., race, gender, ability, and age) to best accomplish the goals of the lesson. • Instructional groups facilitate opportunities for students to set goals, reflect on, and evaluate their learning.	• The instructional grouping arrangements (either whole class, small groups, pairs, individual; heterogeneous or homogenous ability) adequately enhance student understanding and learning efficiency. • Most students in groups know their roles, responsibilities, and group work expectations. • Most students participating in groups are held accountable for group work and individual work. • Instructional group composition is varied (e.g., race, gender, ability, and age) most of the time to best accomplish the goals of the lesson.	• The instructional grouping arrangements (either whole-class, small groups, pairs, individual; heterogeneous or homogenous ability) inhibit student understanding and learning efficiency. • Few students in groups know their roles, responsibilities, and group work expectations. • Few students participating in groups are held accountable for group work and individual work. • Instructional group composition remains unchanged irrespective of the learning and instructional goals of a lesson.
Teacher Content Knowledge ☐	• Teacher displays extensive content knowledge of all the subjects she or he teaches. • Teacher regularly implements a variety of subject-specific instructional strategies to enhance student content knowledge. • The teacher regularly highlights key concepts and ideas and uses them as bases to connect other powerful ideas. • Limited content is taught in sufficient depth to allow for the development of understanding.	• Teacher displays accurate content knowledge of all the subjects he or she teaches. • Teacher sometimes implements subject-specific instructional strategies to enhance student content knowledge. • The teacher sometimes highlights key concepts and ideas and uses them as bases to connect other powerful ideas.	• Teacher displays under-developed content knowledge in several subject areas. • Teacher rarely implements subject-specific instructional strategies to enhance student content knowledge. • Teacher does not understand key concepts and ideas in the discipline and therefore presents content in a disconnected manner.
Teacher Knowledge of Students ☐	• Teacher practices display understanding of each student's anticipated learning difficulties. • Teacher practices regularly incorporate student interests and cultural heritage. • Teacher regularly provides differentiated instructional methods and content to ensure students have the opportunity to master what is being taught.	• Teacher practices display understanding of some student anticipated learning difficulties. • Teacher practices sometimes incorporate student interests and cultural heritage. • Teacher sometimes provides differentiated instructional methods and content to ensure students have the opportunity to master what is being taught.	• Teacher practices demonstrate minimal knowledge of students anticipated learning difficulties. • Teacher practices rarely incorporate student interests or cultural heritage. • Teacher practices demonstrate little differentiation of instructional methods or content.

APPENDIX A

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)

Questioning 	• Teacher questions are varied and high quality, providing a balanced mix of question types:
	  ○ knowledge and comprehension,
	  ○ application and analysis, and
	  ○ creation and evaluation.
	• Questions require students to regularly cite evidence throughout lesson.
	• Questions are consistently purposeful and coherent.
	• A high frequency of questions is asked.
	• Questions are consistently sequenced with attention to the instructional goals.
	• Questions regularly require active responses (e.g., whole class signaling, choral responses, written and shared responses, or group and individual answers).
	• Wait time (3-5 seconds) is consistently provided.
	• The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex.
	• Students generate questions that lead to further inquiry and self directed learning.
	• Questions regularly assess and advance student understanding.
	• When text is involved, majority of questions are text-based.
		• Teacher questions are varied and high quality providing for some, but not all, question types:
		  ○ knowledge and comprehension,
		  ○ application and analysis, and
		  ○ creation and evaluation.
		• Questions usually require students to cite evidence.
		• Questions are usually purposeful and coherent.
		• A moderate frequency of questions asked.
		• Questions are sometimes sequenced with attention to the instructional goals.
		• Questions sometimes require active responses (e.g., whole class signaling, choral responses, or group and individual answers).
		• Wait time is sometimes provided.
		• The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex.
		• When text is involved, majority of questions are text-based.
			• Teacher questions are inconsistent in quality and include few question types:
			  ○ knowledge and comprehension,
			  ○ application and analysis, and
			  ○ creation and evaluation.
			• Questions are random and lack coherence.
			• A low frequency of questions is asked.
			• Questions are rarely sequenced with attention to the instructional goals.
			• Questions rarely require active responses (e.g., whole class signaling, choral responses, or group and individual answers).
			• Wait time is inconsistently provided.
			• The teacher mostly calls on volunteers and high-ability students.

Academic Feedback 	• Oral and written feedback is consistently academically focused, frequent, high quality and references expectations. • Feedback is frequently given during guided practice and homework review. • The teacher circulates to prompt student thinking, assess each student's progress, and provide individual feedback. • Feedback from students is regularly used to monitor and adjust instruction. • Teacher engages students in giving specific and high-quality feedback to one another.	• Oral and written feedback is mostly academically focused, frequent, and mostly high quality. • Feedback is sometimes given during guided practice and homework review. • The teacher circulates during instructional activities to support engagement, and monitor student work. • Feedback from students is sometimes used to monitor and adjust instruction.	• The quality and timeliness of feedback is inconsistent. • Feedback is rarely given during guided practice and homework review. • The teacher circulates during instructional activities but monitors mostly behavior. • Feedback from students is rarely used to monitor or adjust instruction.

APPENDIX A

General Educator Rubric: Instruction

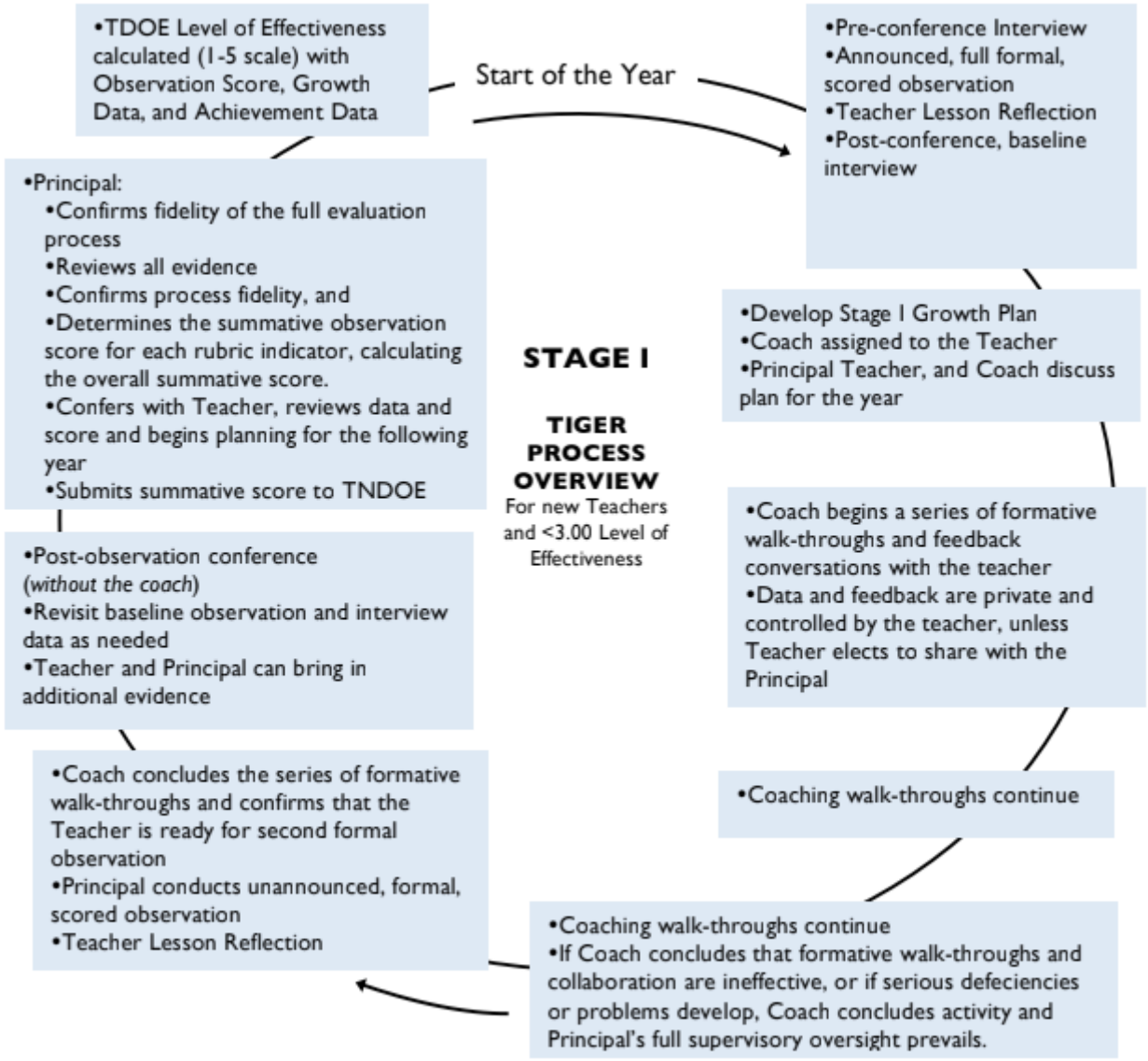
	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)

Thinking 	• The teacher thoroughly teaches two or more types of thinking:
	  ○ analytical thinking, where students analyze, compare and contrast, and evaluate and explain information;
	  ○ practical thinking, where students use, apply, and implement what they learn in real-life scenarios;
	  ○ creative thinking, where students create, design, imagine, and suppose; and
	  ○ research-based thinking, where students explore and review a variety of ideas, models, and solutions to problems.
	• The teacher provides opportunities where students:
	  ○ generate a variety of ideas and alternatives,
	  ○ analyze problems from multiple perspectives and viewpoints, and
	  ○ monitor their thinking to insure that they understand what they are learning, are attending to critical information, and are aware of the learning strategies that they are using and why.
		• The teacher thoroughly teaches one or more types of thinking:
		  ○ analytical thinking, where students analyze, compare and contrast, and evaluate and explain information;
		  ○ practical thinking, where students use, apply, and implement what they learn in real-life scenarios;
		  ○ creative thinking, where students create, design, imagine, and suppose; and
		  ○ research-based thinking, where students explore and review a variety of ideas, models, and solutions to problems.
		• The teacher provides opportunities where students:
		  ○ generate a variety of ideas and alternatives, and
		  ○ analyze problems from multiple perspectives and viewpoints.
			• The teacher implements no learning experiences that thoroughly teach any type of thinking.
			• The teacher provides no opportunities where students:
			  ○ generate a variety of ideas and alternatives, or
			  ○ analyze problems from multiple perspectives and viewpoints.

Problem-Solving 	The teacher implements activities that teach and reinforce three or more of the following problem-solving types: • Abstraction • Categorization • Drawing Conclusions/Justifying Solutions • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing	The teacher implements activities that teach two of the following problem-solving types: • Abstraction • Categorization • Drawing Conclusions/Justifying Solution • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing	The teacher implements no activities that teach the following problem-solving types: • Abstraction • Categorization • Drawing Conclusions/Justifying Solution • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing

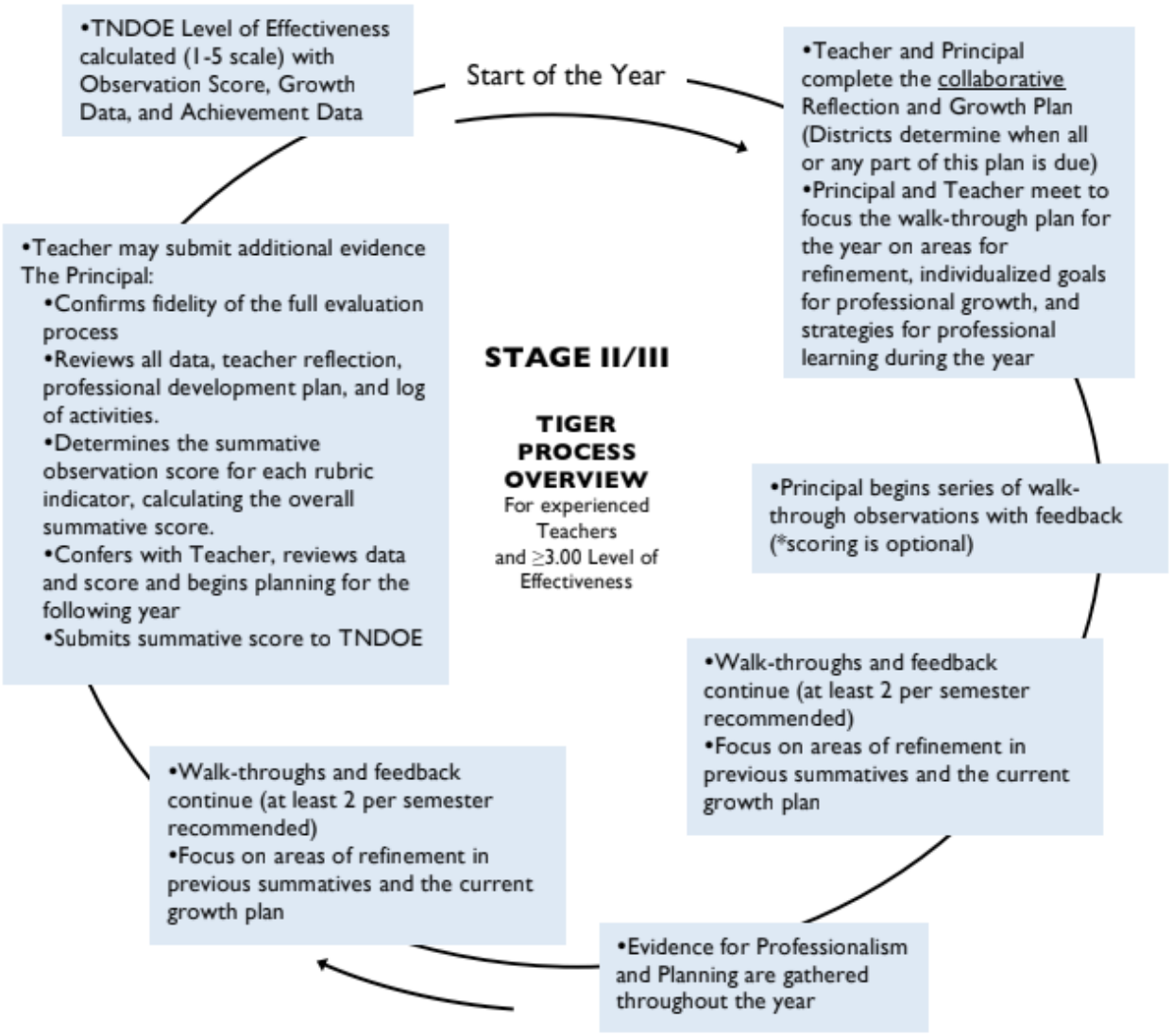
APPENDIX B

TIGER Process Map



APPENDIX B

TEAM Process Map



APPENDIX C

Artifacts Reviewed

GENERAL	District A	District B	District C	District D
• TIGER Process Map • TIGER User Guide, rev. 2017 • TIGER Evaluation Process - A Formtive Model for Teacher Evaluation, Stages & Procedures Handbook • TEAM Evaluation Handbook, 2018 • TEAM General Educator rubric • TIGER/NIET rubric crosswalk • TIGER Districts & TDOE Memorandum of Understanding -TIGER Evaluation Implementation • Tennessee Compass reporting site and electronic feedback platforms	• Observation and feedback form, • Achievement and growth measure selection form, • Walkthrough protocol, • Elementary observation protocol and feedback platform (Google Drive)	• Stage I documents - lesson reflection, pre-conference interview protocol, baseline interview protocol, and growth plan • Stage II documents -- refinements and reinforcements, growth plan • District D observation rubric and teacher self-assesment form	• Content-specific Instructional Practice Guides developed by Achieve the Core and Coaching Stems from Unbound.org, provided	• Observation rubric and teacher self assessment form

APPENDIX D

Interview Protocols

Teachers

Background

How long have you been teaching? When did you first join the faculty at this school?

What attracted you to this district?

What do you currently teach? How did you come to be in this position?

School Context

How would you describe working at this school to another teacher?

What do you see as the primary benefits and challenges of being a teacher here?

Administrative Support and Development

What kinds of supports do you receive to improve your instruction? How well do these supports work for you?

Who do you go to for support related to your teaching? Under what contexts?

Who do you see as the instructional experts in your building?

Teacher Evaluation

Describe the teacher evaluation process in your building.

In your mind, what is the primary purpose of teacher evaluation in this school?

Do you find your evaluation helpful? If so, why? If not, why not?

What aspects of TIGER would you change?

Would you say that teacher evaluation has improved your teaching? Why or why not?

Have you ever been evaluated under another teacher evaluation system? How would you say TIGER processes compare to that experience?

APPENDIX D

Observation & Feedback

How often is your teaching observed?

Who typically observes your teaching?

What happens after your teaching is observed?

What kinds of feedback do you receive? Is it helpful? Why or why not?

How does this relate to the teacher evaluation process?

Would you say that the feedback you receive has improved your teaching? Why or why not?

Teacher Collaboration

Do you meet with other teachers and/or instructional staff to discuss instruction? If so, how frequently?

How would you describe the influence of these meetings on your instruction?

Closing

Is there anything else we should know or you would like to share so that we fully understand this evaluation system and your school/system?

School Administrators (e.g. principals, assistant principals, coaches)

Background

How long have you been in this district? Why did you choose to come here?

What led you to your current role (principal, coach, etc.)?

What did you do before this?

School Context

Tell me about your school leadership team. How is it organized? What roles does it include?

How do you divide responsibilities among the team?

APPENDIX D

Tell me about your teaching force.

- How experienced are your teachers?
- What do you think attracts teachers to this school or district?
- How involved are you in the teacher hiring process (most relevant to principals)?

Support and Development

What supports are available for teachers to improve their instruction?

Are supports different for newer and more experienced teachers?

How would you describe the professional learning structures at this school?

What are the expectations for teachers' engagement in professional learning?

- How are those expectations enforced?

Time Allocation

How do you spend most of your time? (Probe for admin duties, coaching, observation, testing, etc.).

Does this feel like the right allocation to you? If so, why? If not, why not? What do you wish you were doing more/less of?

Teacher Collaboration

Describe the nature of teacher collaboration in your school.

How frequently do teachers and/or instructional staff meet to discuss instruction?

How would you describe the influence of these meetings on classroom instruction?

Teacher Evaluation

Describe the teacher evaluation process in this building.

Who is responsible for what elements of the teacher evaluation process?

In your mind, what is the primary purpose of teacher evaluation in this school?

What do you think works well about the teacher evaluation process in this building? What doesn't work well?

APPENDIX D

What support do you receive in conducting teacher evaluations? (Probe on professional development, central office direction, etc.)

What aspects of TIGER would you change?

Have you ever evaluated teachers under another teacher evaluation system? How would you say TIGER processes compare to that experience?

Observation & Feedback

How frequently are teachers observed in this building? Who conducts the observation?

What happens after teachers are observed?

What kinds of feedback do teachers receive?

How well equipped do you feel to provide feedback to teachers? What gets in the way of providing effective feedback (if relevant)?

How do observation and feedback relate to the teacher evaluation process in your building?

Closing

Is there anything else we should know or you would like to share so that we fully understand this evaluation system and your school/system?

District Staff - (e.g. Central Office - primarily background to understand district policy)

Background

What is your role in the district? How long have you been in this role?

What did you do before this?

Teacher Evaluation

Describe your district's teacher evaluation policy and process.

Describe the AIMS consortium. How do these districts work together? Does AIMS provide any particular development and support around TIGER?

APPENDIX D

Why did your district decide to pursue an alternative observation model? What do you see as the benefits of the TIGER approach? What are some key challenges?

What would you change about TIGER?

Are there particular practices your district implements via its teacher evaluation process that contribute to teacher improvement? What would you say those are?

What supports can principals and/or other evaluators access via the central office? Are these mandatory? If not, how well utilized are they?

How does the district select and support non-evaluative coaches?

What data does the district collect as part of the teacher evaluation process? How is it used?

What does the district see as the primary purpose of teacher evaluation?

Closing

Is there anything else we should know or you would like to share so that we fully understand this evaluation system and your school/system?

APPENDIX E

Qualitative Matrix

Interview Matrix

Position:

District:

	THEMES & EVIDENCE					
Concept	Theme	Key Quotes	Theme	Key Quotes	Documents	Observations
School Concept						
Administrative Support and Development						
Teacher Evaluation						
Observation and Feedback						
Teacher Collaboration						

APPENDIX F

Quantitative Analyses

Table 10. OLS regression of TIGER districts controlling for years of experience, FARMS quartile, and TVASS level.

	2017 Evaluation improves teaching	2018 Evaluation improves teaching	2017 Feedback focused on improvement	2018 Feedback focused on improvement
Years of Experience	-0.0144* (-2.34)	-0.407 (-1.16)	-0.0050* (-2.56)	-0.383* (-2.39)
FARM 0-25%	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
FARM 26-50%	-0.132 (-1.45)	0.0174 (0.23)	-0.0141 (-0.39)	0.0304 (0.89)
FARM 51-75%	0.106 (1.12)	0.123 (1.58)	0.0608 (1.49)	0.0735 (1.80)
FARM 76-100%	0.583 (1.17)	-0.0496 (-0.40)	-0.1648 (-1.34)	0.0618 (1.07)
TVASS Level	-0.0187 (-0.67)	-0.0127 (-0.53)	-0.00534 (-0.47)	0.00255 (0.22)
Intercept	3.396*** (10.04)	3.130*** (17.00)	0.3591*** (6.33)	0.423*** (4.97)
N	359	532	790	954

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

APPENDIX F

Table 11. OLS regression, controlling for case-study districts.

	2017 Evaluation improves teaching	2018 Evaluation improves teaching	2017 Feedback focused on improvement	2018 Feedback focused on improvement
District A	0.0	0.0	0.0	0.0
	(0.0)	(0.0)	(0.0)	(0.0)
District B	-0.080	-0.3267**	-0.114*	0.003
	(-0.78)	(-2.62)	(-1.98)	(0.07)
District C	-0.188	-0.3043**	0.131*	-0.007
	(-1.74)	(-2.34)	(-1.98)	(-0.12)
District D	0.206	0.0944	-0.0420	0.086
	(1.93)	(0.15)	(-0.66)	(1.46)
Intercept	2.942***	3.014***	0.453**	0.294***
	(41.54)	(31.75)	(10.67)	(7.67)
N	299	300	478	513

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

APPENDIX F

Table 12. *Table 12. Target District Analyses:* OLS regression controlling for years of experience, free and reduced priced meal quartile, and TVASS level within case study districts.

	2017 Evaluation improves teaching	2018 Evaluation improves teaching	2017 Feedback focused on improvement	2018 Feedback focused on improvement
Years of Experience	-0.0818*	-0.0967	-0.0500*	-0.0359
	(-2.06)	(-1.26)	(-2.49)	(-1.14)
FARM 0-25%	0.0	0.0	0.0	0.0
	(0.0)	(0.0)	(0.0)	(0.0)
FARM 26-50%	-0.119	0.0502	-0.0873	0.0217
	(-1.18)	(0.33)	(-1.76)	(0.32)
FARM 51-75%	0.109	0.625**	-0.0576	0.0875
	(1.05)	(3.01)	(-1.10)	(0.96)
FARM 76-100%	0.593	0.0785	-0.281*	0.0127
	(1.21)	(0.40)	(-2.12)	(0.15)
TVASS Level	-0.0210	-0.0310	-0.0147	-0.0123
	(-0.74)	(-0.65)	(-1.06)	(-0.59)
Intercept	3.338***	3.153***	0.636***	0.470**
	(15.47)	(8.07)	(5.96)	(2.87)
N	319	147	584	261

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

APPENDIX F

Table 13. *Target District Analyses – 2018 Evaluation Improves Teaching:* OLS regressions for each case study district controlling for years of experience, FARMS quartile, and TVASS level

Evaluation improves teaching	District A	District B	District C	District D
Years of Experience	-0.105 (-1.38)	-0.0919 (-1.18)	0.104 (-1.34)	-0.113 (-1.44)
Free & Reduced Meals Quartile	0.135 (1.90)	0.0510 (0.61)	0.0689 (1.12)	0.113 (1.52)
TVASS level	-0.0139 (-0.30)	-0.0370 (-0.77)	-0.0343 (-0.71)	-0.0389 (-0.80)
High School	0.501* (2.17)	-0.516* (-2.24)	-0.425 (-1.59)	0.129 (0.53)
Middle School	0.461 (1.64)	0.0246 (0.09)	-0.516 (-1.14)	0.518* (2.10)
Elementary School	0.852** (3.19)	-0.306 (-1.07)	-0.524* (-2.47)	-0.133 (-0.49)
Intermediate School	0.302 (1.20)	NA	NA	NA
Intercept	2.808*** (6.98)	3.247*** (7.73)	3.256*** (8.15)	3.084*** (7.68)
N	147	147	147	147

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

APPENDIX F

Table 14. *Target District Analyses - 2018 Feedback is focused on Improvement:* OLS regressions for each case study district controlling for years of experience, FARMS quartile, and TVASS level

Feedback focused on improvement	District A	District B	District C	District D
Years of Experience	-0.0355 (-1.12)	-0.0373 (-1.18)	-0.0360 (-1.15)	-0.0396 (-1.27)
Free & Reduced Meals Quartile	0.0356 (0.11)	0.0106 (0.30)	0.00941 (0.36)	0.0217 (0.70)
TVASS level	-0.0110 (-0.51)	-0.0121 (-0.57)	-0.00824 (-0.39)	-0.0171 (-0.82)
High School	-0.0388 (-0.26)	-0.00828 (-0.08)	-0.0672 (-0.49)	0.0251 (0.24)
Middle School	0.00488 (0.04)	0.0157 (0.13)	-0.0149 (-0.09)	0.305* (2.57)
Elementary School	0.0585 (0.52)	-0.00667 (-0.05)	-0.184 (-1.64)	-0.0558 (-0.47)
Intermediate School	-0.0250 (-0.22)	NA	NA	NA
Intercept	0.481* (2.59)	0.476** (2.69)	0.477** (2.80)	0.461** (2.17)
N	261	261	261	261

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

REFERENCES

Association of Independent and Municipal Schools (AIMS)/TIGER. (2017). AIMS/TIGER teacher evaluation model: Core attributes, procedural components, flexibility options. Copy provided by AIMS district leaders.

Anderson, L.M., Butler, A., Palmiter, A., & Arcaira, E. (November 2016). Study of Emerging Teacher Evaluation Systems. Prepared for: Policy and Program Studies Service, Office of Planning, Evaluation, and Policy Development, U.S. Department of Education. Accessed via: https://www2.ed.gov/rschstat/eval/teaching/emerging-teacher-evaluation/report.pdf

Boyd, D., Grossman, P., Ingo, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). "The influence of school administrators on teacher retention decisions." American Educational Research Journal, 48(2), 303-333.

Bryk, A.S. & Schneider, B. (2002). Trust in schools: A core resource for improvement. New York, NY: Russell Sage Foundation.

Bryk, A.S., Bender Sebring, P., Allensworth, E., Luppescu, S., & Easton, J.Q. (2010). Organizing schools for improvement:
Lessons from Chicago. Chicago, IL: University of Chicago Press.

Bryk, A.S., Gomez, L.M., Grunow, A., & LeMahieu, P.G. (2015). Learning to improve: How America's schools can get better
at getting better. Cambridge, MA: Harvard Education Press.

Campbell, S.L. & Ronfeldt, M. (2018). "Observational evaluation of teachers: Measuring more than we bargained for?" American Educational Research Journal, 55(6), 1233-1267.

Cohen, J. & Goldhaber, D. (2016). "Building a more complete understanding of teacher evaluation using classroom observations." Educational Researcher, 45(6), 378-387.

Coleman, J. (1988). "Social capital in the creation of human capital." American Journal of Sociology, 94: 95-120.

Dee, T.S. & Wyckoff, J. (2015). "Incentives, selection, and teacher performance: Evidence from Impact." Journal of Policy Analysis and Management, 34(2), 267-297.

Donaldson, M. L. (2013). Principals' Approaches to Cultivating Teacher Effectiveness Constraints and Opportunities in Hiring, Assigning, Evaluating, and Developing Teachers. Educational Administration Quarterly, 49(5), 838–882.

Elmore, R.F. (2004). School reform from the inside out: Policy, practice, and performance. Cambridge, MA: Harvard Education Press.

Goldhaber, D. (2016). "In schools, teacher quality matters most." EducationNext, 16(2), 56-62.

Grissom, J. A., Loeb, S., &Master, B. (2013). Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. Educational Researcher, 42(8), 433–444

Grissom, J. & Loeb, S. (2017). "Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low- and high-stakes environments." Education Finance and Policy, 12(3), 369-395.

Grissom, J.A., Rubin, M., Neumerski, C.M., Cannata, M., Drake, T.A., Goldring, E., and Schuermann, P. (2017). "Central offices supports for data-driven talent management decisions: Evidence from the implementation of new systems for measuring teacher effectiveness." Educational Researcher, 46(1), 21-32.

Grissom, J., Bartanen, B., & Jones, A. (2018). "Retaining teachers of color in an era of high-stakes teacher evaluation: Investigating racial differences in teacher evaluation ratings and teacher turnover." Paper prepared for the Association of Public Policy and Management Annual Meeting, November 2018.

Hallinger, P., Murphy, J., & Heck, R.H. (2013). "Leading via teacher evaluation: The case of the missing clothes?" Educational Researcher, 42(6): 349-354.

Hill, H. & Grossman, P. (2013) "Learning from teacher observations: Challenges and opportunities posed by new teacher
evaluation systems." Harvard Educational Review, 83(2), 371-384.

Honig, M. (2010). "District central office leadership as teaching: How central office administrators support principals' development as instructional leaders." Educational Administration Quarterly, 48(4): 733-774.

Howell, W.G. (2015). "Results of President Obama's Race to the Top: Win or lose, states enacted education reforms." EducationNext, 15(4), 58-66.

Hunter, S. (2018). "History of TEAM Teacher Evaluation Policy." Copy provided by author.

Ingersoll, R. (2001). "Teacher turnover and teacher shortages: An organizational analysis." American Educational Research Journal, 38(3), 499-534.

Instruction Partners. (n.d.). "School and district partnerships." Accessed 16 February 2019 via: https://instructionpartners.org/school-district-partnerships

Jackson, K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. American Economic Journal: Applied Economics 1(4):85–108.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. Journal of Labor Economics, 26(1), 101-136.

Kane, T.J. & Staiger, D.O. (2012). "Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains." Bill & Melinda Gates Foundation, Measures of Effective Teaching Project: Research Paper. Accessed via: https://eric.ed.gov/?id=ED540960

Kraft MA, Blazar DL. (2017). Individualized Coaching to Improve Teacher Practice Across Grades and Subjects: New Experimental Evidence. Educational Policy, 31 (7), 1033-1068.

Kraft, M.A., Papay, J.A., Chi, O.L. (2018). Teacher skill development: Evidence from performance ratings by principals. Paper prepared for the Association of Public Policy and Management Annual Meeting, November 2018.

Kraft, M.A. & Papay, J. (2014). "Can Professional Environments in Schools Promote Teacher Development? Explaining

Heterogeneity in Returns to Teaching Experience." Educational Evaluation & Policy Analysis, 36(4), 476-500.

Kraft M.A., Gilmour A. (2016). "Can principals promote teacher development as evaluators? A case study of principals' views and experiences." Educational Administration Quarterly, 52 (5):711-753

Kraft, M.A., Blazar, D., & Hogan, D. (in press). "The effect of coaching on instruction and achievement: A meta-analysis of the causal evidence." Review of Educational Research.

Koedel, C., Li, J., Springer, M., & Tan, L. (2018). "Teacher performance ratings and professional improvement." Journal of Research on Educational Effectiveness, 0(0), 1-26.

Kyndt, E., Gijbels, D., Grosemans, I., & Donche, V. (2016). Teachers' everyday professional development: Mapping informal learning activities, antecedents, and learning outcomes. Review of Educational Research 86(4): 1111-1150.

Leithwood, K., Seashore Louis, K., Anderson, S., & Wahlstrom, K. (2004). How Leadership Influences Student Learning. The Wallace Foundation. Accessed via: https://www.wallacefoundation.org/.../How-Leadership-Influences-Student Learning.pdf

Loeb, S., Darling-Hammond, L., & Luczak, J. (2005). "How teaching conditions predict teacher turnover in California schools." Peabody Journal of Education, 80(3), 44-70.

Murphy, J. & Torre, D.S. (2014). Creating productive cultures in schools. Thousand Oaks, CA: Corwin Press.

National Center for Teacher Quality. (November 2015). "State of the states: Evaluating teaching, leading, and learning." Accessed via: https://www.nctq.org/publications/State-of-the-States-2015:-Evaluating-Teaching,-Leading-and-Learning

Neumerski, C.M., Grissom, J., Goldring, E., Drake, T.A., Rubin, M., Cannata, M., & Schuermann, P. (2018). "Restructuring instructional leadership: How multiple-measure teacher evaluation systems are redefining the role of the school principal." The Elementary School Journal, 119(2), 270-297.

Offices of Research and Education Accountability – Tennessee Comptroller of the Treasury. (March 2012). "Recent teacher policy changes in Tennessee: Teacher evaluation." Accessed via: www.comptroller.tn.gov/Repository/RE Teacher%20Evaluations.pdf

Papay, J. (2012). "Refocusing the debate: Assessing the purposes and tools of teacher evaluation." Harvard Educational
Review, 82(1), 123-141.

Papay J. & Kraft M. (2017). "Developing workplaces where teachers stay, improve, and succeed." In: Teaching in Context: The Social Side of Education Reform. (E. Quintero, Ed.). Cambridge: Harvard Education Press, 15-35.

Papay, J. & Laski, M. (2018). Exploring teacher improvement in Tennessee. Nashville, TN: Tennessee Education Research
Alliance.

Papay, J.P. & Richard, C. (October 2018). Evaluation for teacher development: Exploring the relationship between features of teacher evaluation systems and teacher improvement. Paper prepared for the Association of Public Policy and Management Annual Meeting, November 2018.

Papay, J., Taylor, E. S., Tyler, J., & Laski, M. (2015). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. Unpublished working paper. Available via: https://www.newyorkfed.org/...media/.../education.../pttl-learning-from-colleagues.pdf

Patton, M.Q. (2002). Qualitative research evaluation methods (3rd Ed). Thousand Oaks, CA: Sage Publications.

Peshkin, A. (1993). "The goodness of qualitative research." Educational Researcher, 22(2), 23-29.

Reardon, S.F. (2018). Educational opportunity in early and middle childhood: Variation by place and age. Stanford University: Center for Education Policy Analysis. Accessed via: https://cepa.stanford.edu/content/educational opportunity-early-and-middle-childhood-variation-place-and-age

Rivkin, S.G., Hanushek, E.A., & Kain, J.F. (2005). "Teachers, schools, and academic achievement." Econometrica, 73(2), 417-458.

Roehrich-Patrick, L., Moreo, B., & Gibson, T. (August 2016). "Just how rural or urban are Tennessee's 95 counties: Finding a measure for policymakers." Staff Report to Members of the Tennessee Advisory Commission on Intergovernmental Relations. Accessed via: https://www.tn.gov/content/dam/tn/tacir/documents/2016JustHowRualOrUrban.pdf

Ronfeldt, M. (2012). Where should student teachers learn to teach? Effects of field placement school characteristics on teacher retention and effectiveness. Educational Evaluation and Policy Analysis, 34(1), 3-26.

Ronfeldt, M., Owens Farmer, S., McQueen, K., & Grissom, J. (2015). "Teacher collaboration in instructional teams and student achievement." American Educational Research Journal, 52(3): 475-514.

Shulman, L.S. (1981). "Disciplines of inquiry in education: An overview." Educational Researcher, 10(6), 5-12.

Springer, M., Gegenheimer, K., Hunter, S. & Koedel, C. (Forthcoming). "Does feedback matter: Understanding the relationship between evaluation feedback and teacher productivity." Paper prepared for the Association of Public Policy and Management Annual Meeting, November 2018.

Stecher, B.M., Holtzman, D.J., Garet, M.S., Hamilton, L.S., Engberg, J., Steiner, E.D. … Chambers, J. (2018). "Improving teaching effectiveness: Final report: The Intensive Partnerships for Effective Teaching through 2015-2016." RAND Corporation. Accessed via: https://www.rand.org/pubs/research_reports/RR2242.html

Steinberg, M. & Sartain, L. (2015). "Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching pilot. Education Finance and Policy, 10(4), 535-572.

Sun, M., Mutcheson, R.B. & Kim, J. (2015). "Teachers' Use of Evaluation for Instructional Improvement and School Supports for This Use." In Making the Most of Multiple Measures: The Impacts and Challenges of Implementing Rigorous Teacher Evaluation Systems, eds. Jason A. Grissom and Peter Youngs. New York, NY: Teachers College Press.

Sun, M., Loeb, S., & Grissom, J. (2016). Building Teacher Teams: Evidence of Positive Spillovers from More Effective Colleagues. Education Evaluation and Policy Analysis.

Supovitz, J., Sirinides, P., & May, H. (2010). "How principals and peers influence teaching and learning." Educational Administration Quarterly, 46(1), 31-56.

Tatter, G. (3 February 2016). "Three of Tennessee's four largest districts in the hunt for superintendent." Chalkbeat. Accessed via: https://www.chalkbeat.org/posts/tn/2016/02/03/for-three-of-tennessees-largest-districts simultaneous-superintendent-searches/

Taylor, E.S. & Tyler, J.H. (2012). "The effect of teacher evaluation on teacher performance." The American Economic Review, 102(7), 3628-3651.

Tennessee Department of Education. (n.d.). Tennessee Report Card. Accessed 12 February 2019 via https://reportcard.tnk12.gov/

Tennessee Department of Education. (n.d.). Evaluation, TEAM-TN. Accessed 2 March 2019 via: https://team-tn.org evaluation/

Tennessee Department of Education. (n.d.). Tennessee Academy for School Leaders. Accessed 9 March 2019 via: https://www.tn.gov/education/teaching-in-tennessee/tasl.html

Tennessee Department of Education. (2013). "Teacher Evaluation in Tennessee: A Report on Year Two Implementation."
Accessed via: https://www.tn.gov/content/.../tn/education/reports/rpt_teacher_evaluation_year_2.pdf

Tennessee Department of Education. (2017). "Educator insights: Takeaways from the 2017 Tennessee Educator Survey."
Accessed via: http://educatorsurvey.tnk12.gov/

Tennessee Department of Education. (n.d). "District Profiles, 2017-18." Accessed via: https://www.tn.gov/education/data/data-downloads.html

Tennessee Department of Education. (2018). "TEAM Observation Guidance Documents 2018-19." Accessed via: https://team-tn.org/wp.../08/TEAM-Observation-Guidance-Documents-2018-19.pdf

Tennessee Department of Education. (2018). "TEAM Observation Guidelines 2018-19." Accessed via: https://team-tn.org/wp-content/uploads/2013/.../Observation-Guidelines-2018-19.pdf

Tennessee Department of Education. (2018). "Tennessee Educator Acceleration Model: TEAM Teacher Evaluation: Evaluator Handbook, 2018-19." Accessed via: https://www.tn.gov/content/.../teamtn/TEAM_Teacher_Evaluator Handbook_2018-19.p…

Tennessee Department of Education. (2017). "2018 Tennessee Educator Survey: Statewide Results for Main Survey." Accessed via: http://educatorsurvey.tnk12.gov/#1/all-districts/all-schools/0

Tennessee State Board of Education. (2017, July 28). "Teacher and Administrator Evaluation Policy, 5.201."

Accessed via: https://www.tn.gov/.../tn/stateboardofeducation/.../5.201_Teacher__Principal_Evaluati…

Teacher Instructional Growth and Effectiveness for Results. (n.d.). "TIGER: A formative model for teacher evaluation - About TIGER." Accessed via: http://tigermodel.net/about/

TN Score. (2016). Why more teachers of color should be a priority for Tennessee. [Internet]. Retrieved from: https://tnscore.org/why-more-teachers-of-color-should-be-a-priority-for-tennessee/

Torre, D. & Murphy, J. (2016). "Communities of parental engagement: New foundations for school leaders' work." International journal of leadership in education, 19(2), 203-223.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. TNTP. Accessed via: https://tntp.org/publications/view/the-widget effect-failure-to-act-on-differences-in-teacher-effectiveness

Welch, M., Davis, S., Isaia, R., Johnston, W., Jenuwine, H., & Macdonald, K. (2016). "Aligning evaluation: How much to teacher evaluation rubrics emphasize Common Core instruction?" American Institutes for Research. Accessed via: https://www.air.org/sites/default/files/downloads/report/Teacher-Evaluation-Common-Core Alignment-October-2016.pdf