ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 20-231

Teacher Evaluation, Ambitious Mathematics Instruction, and Mathematical Knowledge for Teaching: Evidence from Early Career

Jihyun Kim
Lehigh University

Ken Frank
Michigan State University

Peter Youngs
University of Virginia

Serena Salloum
Ball State University

Kristen Bieda
Michigan State University

While teacher evaluation policies have been central to efforts to enhance teaching quality over the past decade, little is known about how teachers change their instructional practices in response to such policies. To address this question, this paper drew on classroom observation and survey data to examine how early career teachers' (ECTs') perceptions of pressure associated with teacher evaluation policies seemed to affect their enactment of ambitious mathematics instruction. As part of our analysis, we also considered the role that mathematical knowledge for teaching (MKT) and school norms regarding teaching mathematics shape the potential influence of teacher evaluation policies on ECTs' instructional practices. Understanding how the confluence of these factors is associated with teachers' instruction provides important insights into how to improve teaching quality, which is one of the most important inputs for student learning.

VERSION: May 2020

Suggested citation: Kim, Jihyun, Ken Frank, Peter Youngs, Serena Salloum, and Kristen Bieda. (2020). Teacher Evaluation, Ambitious Mathematics Instruction, and Mathematical Knowledge for Teaching: Evidence from Early Career Teachers. (EdWorkingPaper: 20-231). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/grt9-rg72

Teacher Evaluation, Ambitious Mathematics Instruction, and Mathematical Knowledge for Teaching: Evidence from Early Career Teachers

Jihyun Kim, Ken Frank, Peter Youngs, Serena Salloum, and Kristen Bieda

Over the last decade, teacher evaluation policies have been central to policy efforts to enhance teaching quality. These new systems are expected to enhance teaching quality by providing useful information to teachers about their practices (i.e., formative assessment) as well as imposing sanctions for ineffective teachers (i.e., summative assessment; Hallinger, Heck, & Murphy, 2014). While extant studies have examined how teachers perceived these policies (e.g., Delvaux et al., 2013; Donaldson, 2012; Jiang et al., 2015; Tuytens & Devos, 2009, 2010) and the effects of the policies on student achievement (Dee & Wyckoff, 2015; Steinberg & Sartain, 2015; Taylor & Tyler, 2012), less is known about how teachers change their instructional practices in response to teacher evaluation policies. Such changes are a proximal outcome of the policies that lie between how teachers perceive the policies and the ultimate intended outcome of the policies: improved student achievement. Understanding this part of the policy logic model is essential for evaluating the policies as it sheds light on what actually happens inside of classrooms in response to the policies.

As a step towards addressing this question, we ask how teachers' perceived pressure associated with teacher evaluation policies seems to affect their ambitious mathematics instruction, drawing on unique classroom observation and survey data. Given that "all students should learn and that learning should involve complex ideas and performance" (Lampert, Beasley, Ghousseini, Kazemi, & Franke, 2010, p.129) has become a mantra for educators and policy makers, ambitious mathematics instruction has been emphasized by teacher education

programs as well as the Common Core State Standards (CCSS; Coburn, Hill, & Spillane, 2016; Lampert et al., 2013). Ambitious mathematics instruction aims to improve students' procedural fluency as well as conceptual understanding by providing authentic and meaningful mathematics (Kilpatrick, Swafford & Findell, 2001; Lampert et al., 2010).

However, it is unclear whether teacher evaluation policies motivate teachers to teach mathematics ambitiously. In particular, we focus on early career teachers (ECTs) as the influences of both ambitious instruction and teacher evaluation are likely to be more salient for them. Compared to veteran teachers, ECTs are more likely be trained to teach ambitiously as many teacher preparation programs focus on ambitious instruction and such training might remain fresh in their minds. At the same time, teacher evaluation is designed to have a relatively stronger impact on ECTs' instruction in general (Steinberg & Donaldson, 2016).

Although the CCSS and teacher evaluation policies are both macro-level, external forces imposed on teachers' instructional practices, it is important to take into account the micro-level, internal influence of teachers' own expertise on their instruction. Researchers have placed ECTs' appropriation of knowledge for teaching at the center of understanding their instructional practices (Grossman et al., 2000; Leko & Brownell, 2011; Thompson, Windschitl, & Braaten, 2013). In particular, we focus on ECTs' mathematical knowledge for teaching (MKT), which refers to "a kind of complex mathematical understanding, skill, and fluency used in the work of helping others learn mathematics" (Thames & Ball, 2010, p. 228). While teachers with higher MKT tend to teach more rigorous mathematical content (Hill, Ball, Blunk, Goffney, & Rowan, 2007) and students taught by teachers with higher MKT have significantly higher student achievement gains (Hill, Rowan, & Ball, 2005), it is not clear how such expertise in teaching shapes instruction when teacher evaluation policies come into play. Similarly, we examine how

school norms regarding teaching mathematics shape the potential influence of teacher evaluation policies on ECTs' instructional practices. Specially, we ask the following questions:

1. How are ECTs' perceptions of pressure associated with teacher evaluation policies related to their mathematics instruction?
2. How do ECTs' MKT levels seem to affect the association between their perceptions of pressure related to teacher evaluation policies and their mathematics instruction?
3. How do ECTs' social norms at their school seem to affect the association between their perceptions of pressure related to teacher evaluation policies and their mathematics instruction?

To our knowledge, this study is one of the first to examine the association between pressure from teacher evaluation policies and ECTs' ambitious mathematics instruction. Further, we also take into account other contextual factors, such as teachers' MKT and school norms. Understanding the confluence of those factors for teachers' instruction provides important insights into how to improve teaching quality, which is one of the most important inputs for student learning.

Literature Review

Early Studies in Ambitious Instruction

Ambitious mathematics instruction emphasizes student engagement in problem solving, challenging mathematics tasks, and communication of students' mathematical reasoning. (Cobb, Boufi, McClain, & Whitenack, 1997; Cobb & Smith, 2008). The premise here is that if enacted properly, such instructional practices can promote students' in-depth mathematics knowledge and higher-order thinking through authentic mathematics experience (Lampert et al., 2013).

Indeed, ambitious mathematics instruction has a positive association with student achievement (Blazar, 2015).

The movement towards ambitious instruction was sparked by criticism of the conventional, teacher-centered instruction that was prevalent in the 1980s and 1990s (Brown, Stein, & Forman, 1996), and many teacher preparation programs switched their focus to ambitious instruction accordingly. By definition, however, enacting ambitious instruction is not an easy task, as it involves facilitating complex, unpredictable discussions of student thinking (Ball & Cohen, 1999; Cobb & Smith, 2008; McClain, 2002). Thus, the effects of such pre-service training often remain within the boundary of universities, with ECTs eventually following conventional teaching approaches of their respective schools; this is when the "washing-out effect," "problems of enactment," or "two-worlds" problems occur (Feiman-Nemser & Buchmann, 1985; Stroupe, 2016; Zeichner & Tabachnick, 1981). Once a teacher steps into a classroom, there are many policies and social norms at the school, district, state, and federal levels that he/she has to comply with and such internal and external factors may place pressure on teachers to teach in a certain way that may not be aligned with how they had been trained. For example, if district curriculum is fully scripted and it does not focus on student thinking, ECTs might need to move away from what they learned from their teacher preparation experiences. Thus, to adopt ambitious mathematics instruction is hard for ECTs without sufficient external support.

One source of such support are district-level policies and professional development (Ball & Cohen, 1999; Franke, Carpenter, Levi, & Fennema, 2001; Jennings & Spillane, 1996; Spillane & Jennings, 1997). However, findings from studies on such support are mixed. For instance, Ball and Cohen (1999) pointed out that most professional development sessions were "intellectually

superficial, disconnected from deep issues of curriculum and learning…" and fail to contribute to teachers' learning (p. 4). On the other hand, Garet and colleagues (2001) showed that some professional development programs with a clear focus on particular teaching practices and area knowledge and that had a high level of coherence were effective in improving teaching quality.

Social norms can be inhibitive, but they also can be supportive for ambitious instruction. Several research studies have established clear effects of school-based communities of practice, norms, and/or social networks among teachers on ambitious mathematics instruction (Cobb & Smith, 2008; Coburn & Russell, 2008; Lave, 1996; Rogoff, 1994; Stroupe, 2016). That is, when school norms are aligned with ambitious instruction, it might be easier for ECTs to teach ambitiously since social norms and their training both reinforce each other for ambitious instruction. Following the enactment of accountability policies in the early-2000s, the question became how such policies shape teachers' ambitious instruction.

Teacher Evaluation Policies and Teachers' Instructional Practices

Few studies have examined how teacher evaluation policies shape teachers' ambitious instructional practices. Instead, previous studies mostly depended on indirect or less accurate measures to describe the influence of the policy on instruction. First, some studies evaluated the policies based on student test scores and/or teachers' classroom observation ratings. For example, the implementation of teacher evaluation had positive effects in Cincinnati (Taylor & Tyler, 2012), Chicago (Steinberg & Sartain, 2015), and Washington D.C. (Dee & Wyckoff, 2015). In contrast, a recent report by Stecher and colleagues (2018) showed that teacher evaluation had a null to negative effect on student achievement in three school districts and four charter management organizations in Florida, Tennessee, and Pennsylvania. While these studies provide important insights into the effects of teacher evaluation policies, the outcomes (i.e.,

student test scores and classroom observation ratings) for these studies were also part of summative teacher evaluation which could be used for making high-stakes decisions about teachers' job status. Thus, using these measures might introduce some validity threats. In order to avoid these potential biases, it is imperative to analyze external observers' ratings free from any high-stakes decisions to understand the potential influence of teacher evaluation on teachers' instructional practices. More importantly, the observation rubrics that previous studies used mostly focus on general pedagogy, rather than measuring ambitious instruction for a specific subject.

Second, some studies on teacher evaluation policies have relied on teachers' perceptions of the effects of policies on their instructional practices (e.g., Delvaux et al., 2013; Donaldson, 2012; Donaldson, Woulfin, LeChasseur, & Cobb, 2016; Koedel, Li, Springer, & Tan, 2019; Milanowski & Heneman, 2001; Tuytens & Devos, 2010). For example, Donaldson (2012) found that most teachers reported that teacher evaluation policies shaped how they planned lessons and their general approach to teaching, while the feedback they received from teacher evaluation did not affect their instructional practice. Delvaux and colleagues (2013) reported that some teachers found that teacher evaluation was useful for professional development, which in turn, could have affected their instruction. In contrast, Koedel et al. (2019) found no effect of teacher evaluation ratings on teachers' self-reported professional improvement activities.

To be sure, the perceptions of those who enact a policy are a critical factor in the success of that policy (Bridwell-Mitchell & Sherer, 2017; Coburn, 2001; Spillane, Reiser, & Reimer, 2002) and teachers' perceptions of teacher evaluation policies have been used in numerous studies to examine the implementation of the policies (e.g., Jiang et al., 2015; Kim, Sun, & Youngs, 2019; Kraft & Gilmour, 2017; Reinhorn, Johnson, & Simon, 2017). However,

measuring the effects of policies by solely relying on teachers' reports about their behaviors might introduce measurement error, and doing so misses an important piece of information about teachers' behavioral responses to the policies as teachers might not be aware of significant changes in their behavior.

In sum, previous literature on teacher evaluation policies suggests that these policies are likely to shape teachers' instructional practices but the positive or negative direction of the influence is still unclear. To address this gap in the literature, it is clear that we need accurate data on teachers' instructional practices which are not attached to high-stakes decisions and beyond what teachers think about their own practices.

Theoretical Framework

We first drew on a framework developed by Frank et al. (forthcoming) to conceptualize various factors that affect ECTs' planning and enactment of ambitious instruction (see Figure 1).

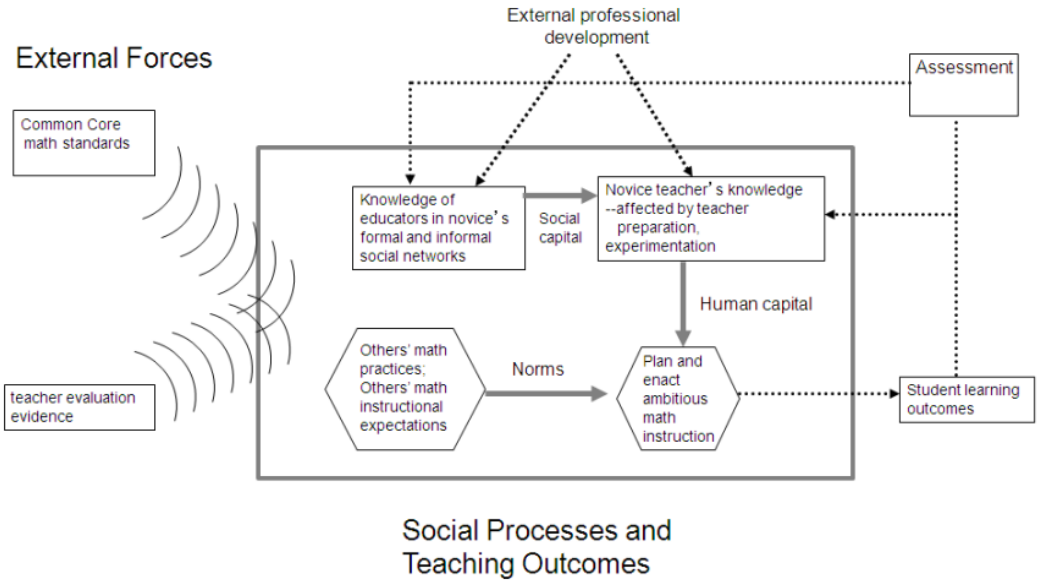

Figure 1. Factors that Influence Novice Teachers' Knowledge and Math Instruction

There are four primary factors that shape teachers' instructional practices according to this conceptualization: 1) external forces, such as the CCSS and teacher evaluation; 2) social capital and norms within schools; 3) teachers' own human capital, such as their MKT; and 4) feedback and external support including student assessment and professional development. In the current study, we modified this broader framework to focus more on the potential influence of teacher evaluation on ECTs' ambitious mathematics instruction (see Figure 2).

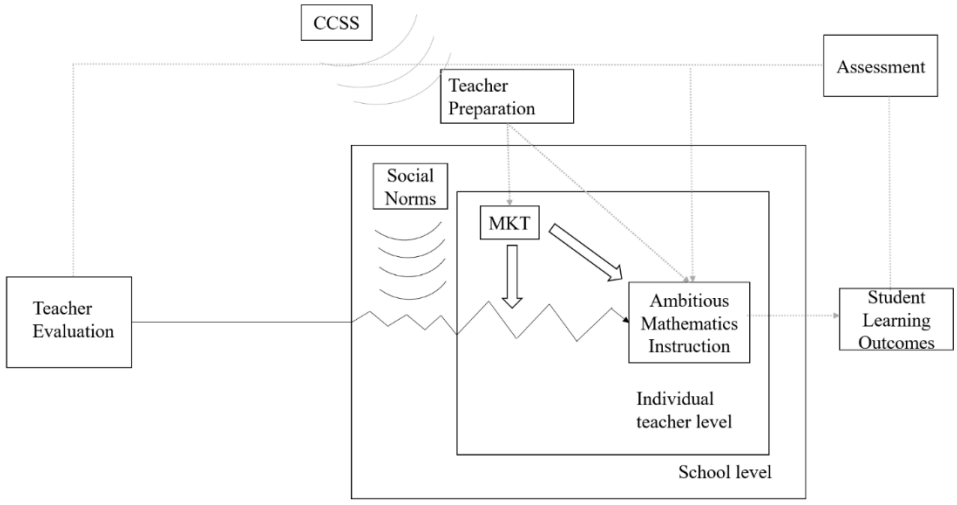

Figure 2. The Influence of Teacher evaluation on Ambitious Mathematics Instruction

Teachers' instructional practices are situated in a larger institutional setting (Cobb & Smith, 2008), and teacher evaluation policies that define effective teaching for teachers are likely to affect teaching practices. In this process, however, teachers often "adapt, adopt, combine or ignore (policy) messages and pressures…" (Coburn, 2001, p. 147). Such individual and collective sensemaking are what prompt teachers to act in response to their perceptions of a policy, rather than the written policy itself. Thus, we use teachers' perceived pressure of teacher evaluation policies as a proxy in order to operationalize the impact of such policies. To be sure,

self-reported perceptions of pressure might not represent all of the ways that the policies affect teachers. The ideal situation would be a randomized control trial with a treatment group of teachers who were evaluated and a control group of teachers who were not evaluated. However, this would only be possible when teacher evaluation policies were applied to only a subset of teachers; e.g., in a pilot process. Under circumstances where almost all teachers are required to be evaluated, this approach is not feasible. Moreover, being exempted from the annual evaluation process is very rare among ECTs.

Instead, we focus on teachers' perceptions of the strength of the pressure of teacher evaluation, and how this perception is associated with teachers' actual teaching practices. We acknowledge that teachers' perceived pressure might not be the same as the actual pressure associated with teacher evaluation. However, we argue that it is how teachers make sense of policies that leads to changes in their behaviors, rather than the actual policy. In Figure 2, the line from teacher evaluation to ambitious mathematics instruction captures this phenomenon of individual and collective sensemaking of the policies. When the line meets school- and individual-level contexts, it becomes non-linear as a result of teachers' sensemaking processes.

At the individual-teacher level, we examine how teachers' MKT shapes the influence of teacher evaluation on their instructional practices. While MKT is an active ingredient of high-quality mathematics teaching, it should be noted that having a high level of MKT does not guarantee that a teacher will engage in ambitious mathematics instruction, as MKT does not target pedagogical quality itself, but instead knowledge that teachers need to teach mathematics in a rigorous way (Hill et al., 2005; Hill et al., 2007). Thus, we conceptualize teachers' MKT as their capacity to enact ambitious instruction; a teacher with a very high level of MKT has the

potential to reach the high end of ambitious mathematics teaching. However, at the same time, a teacher with high MKT can teach at a low level if they choose to focus primarily on basic skills.

In contrast, a teacher with a very low level of MKT is more likely to stay at the lower end of ambitious mathematics teaching, as they do not possess enough knowledge to enact such instruction. More importantly, we conceive that the bound on ambitious instruction placed by MKT may interact with evaluation pressure. Teachers with low levels of MKT may not alter their teaching practices in response to evaluation pressure because their practices are already limited by their MKT. In contrast, teachers with high MKT may experience evaluation pressure as a constraint on the range of ambitious instruction in which they can engage. On the other hand, it should be noted that the same analysis could test how the impact of MKT on teacher instruction changes in response to teacher evaluation. We take into account the nature of this two-way interaction term when we interpret our results.

At the school level, we focus on social norms in schools. Strong support from teacher networks is imperative for ambitious instruction (Cobb & Smith, 2008; Coburn & Russell, 2008; Lave, 1996; Rogoff, 1994; Stroupe, 2016). Especially for ECTs, other teachers' enactment of ambitious mathematics instruction in their respective classes, which contributes to school norms regarding mathematics instruction, is an important resource. That is, social norms at a given school can accelerate the penetration of external pressure on teachers, or they can filter out such influence (Frank et al., 2011; Penuel, Sun, Frank, & Gallagher, 2012; Sun, Frank, Pennuel, & Kim, 2013). Thus, we hypothesize that the influence of teacher evaluation on ECTs' ambitious mathematics instruction can differ based on varying school social norms.

In addition, we acknowledge other factors that can shape teachers' instructional practices. The CCSS and other equivalent state-level mathematics content standards can also shape

teachers' instructional practices directly and/or indirectly through teacher preparation and other initiatives. Under the current system of teacher evaluation where student achievement scores typically receive substantial weight, there is a feedback loop between student learning outcomes and the influence of teacher evaluation. While these parts of the framework are critical for understanding the context of teachers' efforts to enact ambitious mathematics instruction, they are beyond the scope of the current paper, so the arrows among those parts of Figure 2 appear in gray.

Method

Sample

In the 2014-15, 2015-16, and 2016-17 school years, as part of a larger research project we contacted all early career teachers in grades K-6 who had up to four years of full-time experience in elementary schools in eight school districts in three Midwestern states. The current study reported here focuses only on data collected in the 2015-16 school year, due to the discrepancies in survey instruments across years. Among these three states, two states implemented the CCSS, and the other state implemented a version of state mathematics standards that was very similar to the CCSS as of 2015-16. Table 1 reports background information for the sampled districts.

Most districts in the study had different teacher evaluation systems for tenured- and non-tenured teachers, and the systems for non-tenured teachers generally featured more frequent evaluations. In addition, the results from the evaluation of non-tenured teachers were used as a critical source of information for determining their employment status for the following school year. Although there was some variation in evaluation components across districts, most systems included student growth measures calculated at different levels (classroom-, building-, and/or district-level growth) and observations based on various tools, such as Five Dimensions of Teaching and Learning (Center for Educational Leadership, n.d.), Danielson's Framework for

Teaching (Danielson, 2013), or a rubric developed by the district. Most districts used students' standardized tests scores for calculating student growth measures, but some districts deferred the student growth component altogether to the 2016-17 school year. Teachers were evaluated with regard to four or five levels of performance based on the weighted total score of each component of evaluation. Many districts in the study put the largest weights on observations conducted by building-level administrators.

Eligible ECTs were asked to complete three surveys during 2015-16, including an MKT survey and two surveys that asked about their perceptions of teacher evaluation and their own social networks in their schools. They were also asked to participate in four observations of their mathematics lessons during the 2015-16 school year, two in the fall and two in the spring. The observations included both video ratings and live observations. A total of 296 ECTs were contacted for observations and surveys, and 102 of them participated the study (a 34% participation rate). Such a rate is not ideal, but given that we asked ECTs to participate in four observations and three surveys in a single school year, it is not particularly low. As Frank et al. (forthcoming) discussed, compared to other studies on teachers' instructional practices and school contextual factors (e.g., Smith, Booker, Hochberg, & Desimone, 2018), this is a relatively large data set that includes both classroom observations and information about policy and school contexts.[1]

Table 1. Background Information on Participating Districts in 2015-16

District	State	Number of K-12 students	% of free/reduced lunch eligible students	% of White students
District A	State 1	19,000	70%	92%
District B	State 1	10,000	39%	91%
District C	State 2	6,000	5.3%	86%
District D	State 2	17,000	63%	28%
District E	State 3	8,000	23%	79%
District F	State 3	21,000	14%	77%
District G	State 3	15,000	61%	36%
District H	State 3	12,000	72%	32%

Note. The number of K-12 students in each district is rounded for de-identification of school districts. Source: Common Core of Data (National Center for Education Statistics).

Figure 3 presents the timeline of data collection. In the first survey, administered in fall 2015, ECTs were asked to list their close teacher colleagues and formal mentor. Nominated teachers were then contacted and asked to complete two surveys, including an MKT survey and a survey about their enactment of mathematics instruction. The data collected from this process is egocentric network data because we only focused on the social networks of ECTs. This contrasts with socio-centric network data, which would include data on the social networks of all teachers in a given school. We contacted 282 mentors and colleagues nominated by participating ECTs and asked them to complete two surveys in winter 2016. A total of 158 teachers completed both surveys (a 56% response rate). In the next semester, spring 2016, we also asked ECTs to list their colleagues with whom they had discussed mathematics instruction in the past six months and their mentor. We used this spring social network data and close colleagues' self-reported instructional practice in winter 2016 to calculate their exposure to social norms, which will be explained in the following section.

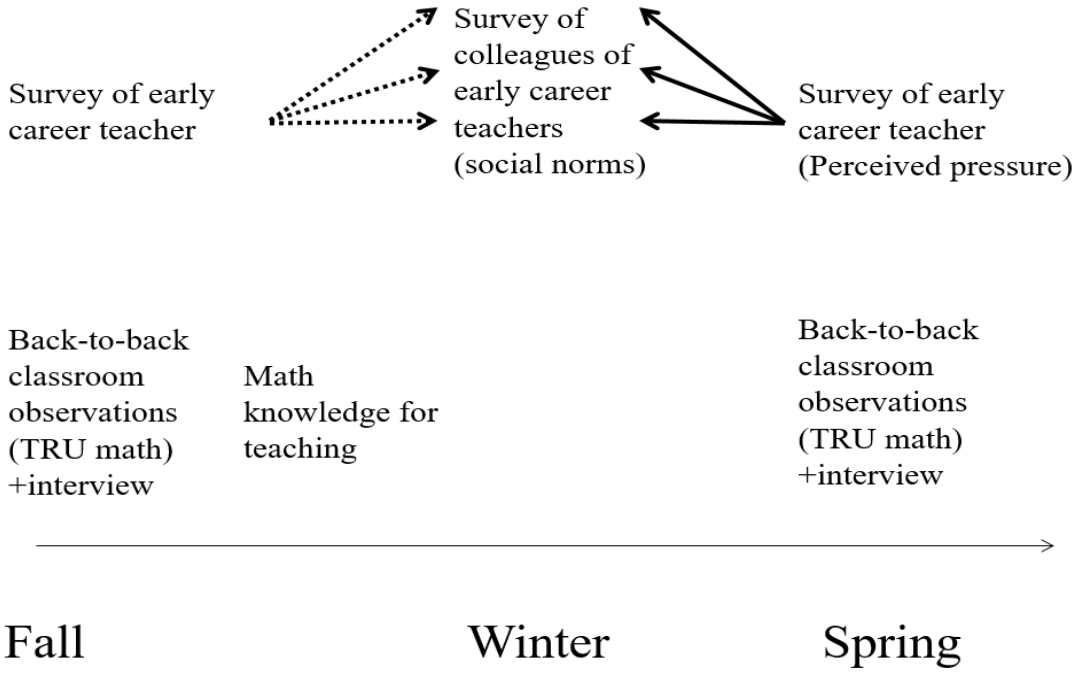

Figure 3. Data Collection Timeline.

Measures

ECTs' enactment of mathematics instruction (TRU Math). Members of the research team carried out two back-to-back observations of each ECT in fall 2015 and again in spring 2016 to measure the quality of their mathematics teaching by using the TRU Math observation tool (Teaching for Robust Understanding in Mathematics; Schoenfeld, Floden, & the Algebra Teaching Study and Mathematics Assessment Project, 2014). The purpose of this subject-specific tool is to define and measure classroom interactions that enhance students' "robust understanding" of mathematical concepts (Schoenfeld, 2013, p. 608). We used a modified version of the TRU Math rubric consisting of four dimensions: 1) the mathematics; 2) cognitive demand; 3) agency, authority, and identity; and 4) uses of assessment (See Appendix A for The

TRU Math Scoring Rubric). The TRU Math rubric is grounded in the literature on the features of ambitious mathematics instruction (See Schoenfeld, 2014 for detailed discussion of theoretical underpinning of TRU Math). To be specific, dimensions 1 and 2 measure whether the mathematics discussed in the class is focused, coherent, and challenging, making meaningful connection between procedure and concepts (Schoenfeld, 2014). Note that connection between procedural and conceptual understanding is one of the main goals for ambitious instruction. Dimensions 3 and 4 attend to how teachers surface and use students' ideas to deepen the conversation, which is another important feature of meaningful and authentic mathematics instruction.

Raters coded each lesson in 10-minute units for each dimension, according to the rubric. The scale was from one to three, treating non-mathematical activities as missing values and allowing .5 scales. It should be noted that the rubric is specified for each activity type, including whole class, individual work, and small group. In order to maintain a high level of inter-rater reliability (IRR), all raters had regular meetings to discuss the TRU Math observation rubric and specific cases. The whole team of raters maintained a weighted kappa threshold of 0.5 throughout the process of coding episodes which is an acceptable level of reliability of observation ratings (Bell et al., 2014). The details of how the project selected the tool and trained raters are documented in Bieda et al. (forthcoming).

The episode-level data for the four dimensions collected in spring 2016 were used as the main dependent variable for the analysis. Fall 2015 data were used as one of the main control variables in order to account for unobservable factors that might have biased the analysis. For example, student characteristics and school leadership can affect both teachers' perceived pressure related to teacher evaluation and mathematics instruction. Since those factors might

have already affected teachers' instruction in fall 2015, including this pre-measure helped reduce potential bias (Cook, Shadish, & Wong, 2008). The average scores in fall 2015 for each dimension were included at the teacher level as the pre-measure.

ECTs' perceived pressure related to teacher evaluation. In the spring 2016 ECT survey, several items about teachers' perceptions of teacher evaluation policies were included: 1) "The current teacher evaluation system has significantly affected my mathematics instruction;" 2) "I need to change my current teaching practices in order to earn a high score;" 3) "I need to earn a high teacher evaluation score to keep my job;" and 4) "I am concerned that my evaluation results can be used in making decisions" (α=0.642). ECTs were asked about the extent to which they agreed with these statements (strongly disagree=1, disagree=2, agree=3, and strongly agree=4). As noted above, these items measured the magnitude of ECTs' perceptions that policies were contributing to changes in their instruction. Two survey items on the influence of teacher evaluation on instructional practices are similar to ones in the UChicago Consortium on School Research's 5 Essentials for School Improvement survey (2017), which has been used for several extant studies on teacher evaluation (e.g., Jiang et al., 2015; Jiang & Sporte, 2016). The last two items were created to focus on teachers' job status as we focus on ECTs. We took the mean of ECTs' responses to these four items and used it as the main independent variable for the analysis.

ECTs' MKT. In order to measure the level of teachers' MKT, we administered a survey developed by Hill and colleagues (2004). This instrument measures teachers' knowledge for teaching various mathematical topics and different domains of teacher knowledge, such as knowledge of students and content. As noted earlier, MKT has long been regarded as a valid measure of teachers' knowledge for teaching (Hill et al., 2005; Hill et al., 2007). The MKT

survey used for the project focused on elementary number and operation concepts, and scores were provided as IRT scores by University of Michigan's Learning Mathematics for Teaching Project online system.

Social norms. In order to capture social norms regarding mathematics teaching, we asked ECTs in the fall and spring surveys to list up to 10 teachers with whom they discussed mathematics instruction within the past six months at the same school. Then we also asked about the frequency of their interactions with these individuals (1 = less than once a month; 2 = 1 to 3 times a month; 3 = 1 to 2 times per week; 4 = 3 to 4 times per week; 5 = every day). The nominated teachers whom we contacted answered questions about their enactment of mathematics instruction. The stem asked, "During the last 5 math lessons that you taught, in how many did your students have opportunities to do each of the following?" and the items included: 1) "Verbally express their thinking;" 2) "Make connections between different strategies;" and 3) "Discuss other students' strategies." These three items align well with some elements of ambitious instruction manifested in the CCSS (Welch et al., 2016). Teachers answered based on a scale of 0 lessons, 1 to 2 lessons, 3 to 4 lessons, and 5 lessons (α=0.791), and we took a standard mean of their responses to these three survey items.

Although we contacted ECTs' social network members based on fall nominations, we used spring social network data to calculate social network exposure to school norms. That is, only the network members who had been nominated in fall and spring by at least one ECT at their school were included in the analysis.[2] Despite the risk of losing data, we adopted this approach to best capture the average level of social norms at schools that ECTs might experience between the fall and spring semesters. It should be noted that we asked ECTs about their social networks in the past six months in the spring survey and collected nominated teachers'

instructional practices in the winter, two to three months before the spring ECT survey was distributed.

Following the approach of Sun et al. (2013), social network exposure to school norms was calculated by taking the mean of the instructional practices of nominated colleagues weighted by the frequency of interaction between the ECT and these individuals. Social network exposure to school norms is specified as follows (Sun et al., 2013):

$$\text{Exposure to SocialNorms}_i = 1/n_i \sum_{i'=1}^{n} (Interaction_{ii'}) \times (\text{Colleagues' practice}_{i'}) \quad (1)$$

Where n_i is the total number of teachers nominated in spring by ECT *i* whose instructional practice data were available.

Covariates. In addition, various control variables were included in the analysis in order to improve the precision of the model. At the teacher level, whether a teacher taught a grade tested by state-standardized tests (i.e., 3rd- to 8th-graders for all three states in the current study), the number of students in the class, whether the observation was video or live, whether a teacher held a master's degree, years of teaching experience and the intensity of professional development related to mathematics instruction were included. The professional development variable was measured based on the ECTs' response to one survey item in spring: During the past 6 months, how many hours/days did you spend addressing mathematics instruction in school or district induction or professional development activities? (none=1; 1-2 hours=2; 3-4 hours=3; 5-8 hours=4; 2 days=5; 3 or more days=6). As noted above, however, the most important control variable is the teacher's pre-measure of ambitious instruction based on observations at the beginning of the 2015-16 school year.

Along with these control variables, district fixed-effects are included for all analyses in order to control for various attributes of districts that affect both teachers' perceptions of teacher evaluation pressure and their instructional practices, such as curriculum and student composition.

Missing values. Following Cohen, West and Aiken (2014), the missing values for evaluation pressure, the intensity of professional development, whether a teacher held a master's degree, teachers' years of teaching, and social norms were imputed in models. Up to 14% of teachers had missing values on the first four variables respectively and about 30% of teachers missed the social norms variable. Whether a teacher held a master's degree variable was imputed as 0 when it is missed (i.e., not holding a master's degree) and other variables were imputed by grand-mean. We included indicators for missing cases for these variables in the model and we ran the models with and without these imputed variables (See Appendix C for results without imputation).

Analytical Approach

Document analysis. Before we analyzed the data collected from teachers in our study, we investigated one of the teacher evaluation rubrics used by a few sample districts according to the protocol developed by Welch and colleagues (2016) to understand how much the teacher evaluation rubric overlapped with the CCSS (See Appendix B for details). Although not all aspects of ambitious mathematics instruction were manifested in the CCSS, analyzing the alignment between the CCSS and this teacher evaluation rubric helped us to understand the policy messages that teachers received from these two external forces. We selected a teacher evaluation rubric that principals used for classroom observations because in all districts, classroom observations were given the largest weight among different teacher evaluation components. We focused on evaluation rubrics in one state that contained half of our

participating districts in the district sample and more than half of our participating teachers in the teacher sample. Those districts modified the state evaluation rubric, but the core indicators remained very similar. The teacher evaluation rubric includes planning, instruction, and teacher leadership. The rubric was coded by two different co-authors initially and discrepancies in the coding were discussed until the two coders reached consensus.[3]

Quantitative analysis. For the quantitative analysis on the association between evaluation pressure and ambitious instruction, it was important to determine the level of analysis for each research question. Based on the research design, there could potentially be four different levels: episode-level TRU Math scores are nested in teachers, teachers are nested in schools, and schools are nested in districts. In order to take the nested structure of the data into account, three- or four-level hierarchical linear modeling (HLM) may be considered theoretically (Raudenbush & Bryk, 2002). However, the school level was excluded because there were not enough ECTs per school; on average, there were slightly less than two ECTs per school in the data. We also ran an unconditional HLM model of teachers within districts to assess how much variation in perceived pressure was at the teacher versus district level. Since districts have slightly different systems for evaluation, there may be some variance in teacher perception across districts. The Intra Class Correlation (ICC) was almost zero at the district level, so we concluded that teachers' perceived policy pressure varied primarily within districts, which justifies our use of district fixed effects. Since teachers have different number of episodes and episode level data are nested at the teacher level, we decided to include episode level in the model. Taken together, we applied two-level HLM for our analysis where the first level is the episode level and the second level is the teacher level with district fixed effects. As a robustness check, we also run an OLS regression with the teacher level aggregated data (See Appendix C for results).

The main model for these research questions is specified as follows:

Level 1(Episode level): $Y_{ij} = \beta_{0j} + e_{ij}$

Level 2(Teacher level): $\beta_{0j} = \gamma_{00} + \rho(\text{Pre-measure}_j) + \gamma_{01}(\text{Perceived Evaluation Pressure}_j)$

$$+ \gamma_{02}(\text{MKT}_j) + \gamma_{03}(\text{SocialNorms}_j) + \mathbf{Z}_j\lambda + \mathbf{D}_j\eta + u_{0j} \quad (2)$$

Where Y_{ij} indicates a TRU Math score for episode *i* for teacher *j.* Scores for the four TRU Math dimensions are included separately and one model includes the average of all dimensions. Pre-measure_j is an average score of each dimension in fall 2015. For example, when the outcome is dimension 1 (the mathematics) score of an episode in spring observations, Pre-measure_j is the average score of dimension 1 in the fall observations across all episodes for the same teacher, so it is included at the teacher level. $\text{Perceived Evaluation Pressure}_j$ is ECT *j*'s perceived pressure associated with teacher evaluation; MKT_j and SocialNorms_j are their MKT test scores and their colleague's enactment of ambitious instruction respectively. The term $\mathbf{Z}_j$ represents a vector of covariates such as whether a teacher taught a grade tested by state-standardized tests (i.e., 3rd- to 8th-graders for all three states in the current study), whether the observation was video or live, whether a teacher held a master's degree, years of teaching experience, the intensity of professional development, and the number of students in the class. $\mathbf{D}_j$ are district-fixed effects, and e_{ij} and u_{0j} are episode specific- and teacher specific-residuals, respectively. As noted above, school level was excluded as there are not enough number of teachers per school, but we also run a school fixed effects model with a teacher-level analysis as a secondary analysis (See Appendix C).

For the research question about potential moderating the effects of ECTs' MKT, a term for the interaction between ECTs' MKT and the perceived pressure variable is added to model

(2). The variables for perceived evaluation pressure and MKT are grand-mean centered for ease of interpretation and to account for multicollinearity.

Level 1: $Y_{ij=} \beta_{0j} + e_{ij}$

Level 2: $\beta_{0j=}\gamma_{00}+ \rho(\text{Pre-measure}_j)+ \gamma_{01}(\text{Perceived Evaluation Pressure}_j)+ \gamma_{02}(\text{MKT}_j)+$

$\gamma_{03}(\text{SocialNorms}_j)+\gamma_{04}(\text{MKT}_j\text{* Perceived Evaluation Pressure}_j) +\mathbf{Z}_j\lambda+\mathbf{D}_j\eta+u_{0j}$ (3)

The next question is related to the moderating effects of social norms on the association between the two main variables. In this model, we included the social network exposure term calculated by equation (1).

Level 1: $Y_{ij=} \beta_{0j} + \mathbf{X}_{ij}\delta + e_{ij}$

Level 2: $\beta_{0j=}\gamma_{00}+ \rho(\text{Pre-measure}_j)+ \gamma_{01}(\text{Perceived Evaluation Pressure}_j)+ \gamma_{02}(\text{MKT}_j)$

$+\gamma_{03}(\text{SocialNorm}_j)+\gamma_{04}(\text{SocialNorms* Perceived Evaluation Pressure}_j)+\mathbf{Z}_j\lambda+\mathbf{D}_j\eta+u_{0j}$ (4)

Results

Document Analysis

Before we turn to our quantitative analysis, we analyze a teacher evaluation rubric that was frequently used by our sample districts to understand whether there was significant overlap or conflict between CCSS and teacher evaluation by design. This document analysis is the first step to understand potential tensions between ambitious instruction and teacher evaluation as it provides some evidence at the policy level.

Out of 18 dimensions, only 5 dimensions were coded as "present," which means that CCSS criteria appeared in the evaluation rubric (27.78% alignment, see Appendix B for details). This is close to the average alignment scores between CCSS and state models (33%) in Welch et al. (2016). This result suggests that the policy messages that teachers received about the CCSS and teacher evaluation might not be well-aligned with one another. In other words, teachers

might not be able to satisfy expectations related to both teacher evaluation and ambitious instruction; they were likely to be forced to choose one or another.

Perceived Pressure Related to Teacher Evaluation and Ambitious Mathematics Instruction

Table 2 reports the descriptive statistics of the main variables for the analysis. ECTs generally scored higher in TRU Math dimensions 1 and 2 than in dimensions 3 and 4. This pattern in TRU Math ratings indicates that ECTs were relatively better prepared to plan and enact ambitious tasks while their abilities to facilitate student discussion might have not been fully developed yet. In addition, in spring dimensions 1 and 2 ratings, and dimensions 3 and 4 ratings are highly correlated with one another ($r=0.80$, p-value≤ 0.001 and $r=0.84$, p-value≤ 0.001, respectively). This finding indicates that although theoretically each dimension of TRU Math measured different aspects of ambitious instruction, empirically there might be two main aspects of mathematics teaching measured by this tool. For instance, dimensions 1 and 2 captured some aspects of the tasks discussed during the class, whereas dimensions 3 and 4 measured the quality of interaction between students and teachers.

Table 2. Descriptive Statistics for Key Variables

Variable	M	SD	Min	Max	N
Episode level characteristics in spring 2016 (Outcome variable)					
The mathematics	1.966	0.513	1	3	1188
Cognitive demand	1.741	0.546	1	3	1177
Agency, authority, and identity	1.454	0.492	1	3	1169
Uses of assessment	1.487	0.539	1	3	1172
Four Dimensions Combined	1.656	0.442	1	3	1212
Teacher level TRU Math in fall 2015 (Covariate)					
The mathematics	1.939	0.410	1.083	2.917	96
Cognitive demand	1.684	0.400	1	2.75	96
Agency, authority, and identity	1.424	0.320	1	2.5	96
Uses of assessment	1.504	0.379	1	2.714	96
Four Dimensions Combined	1.638	0.342	1.030	2.612	96
ECTs' characteristics					
Perceived evaluation pressure	2.153	0.523	1	3.75	90
Missing flag_evaluation pressure	0.143	-	0	1	105
MKT	-0.082	0.924	-2.007	2.513	95
Teaching tested grade	0.314	-	0	1	102
Holding a Master's degree	0.255	-	0	1	94
Missing flag_ a Master's degree	0.105	-	0	1	105
Total years of experience working as a certified teacher	2.711	1.222	0	5	90
Missing flag_ total years of experience	0.143	-	0	1	105
Professional development	3.663	1.507	1	6	92
Missing flag_ professional development	0.124	-	0	1	105
The number of students in the class	24.892	4.791	18	36	102
Video observation	0.594	-	0	1	96
Social network members' characteristics					
Social network members' enactment of instruction	7.771	2.899	3	15	74
Missing flag_ Social network members' enactment of instruction	0.295	-	0	1	105

Note. Video observation, teaching tested-grade, holding a Master's degree, all missing flags are dummy variables.

In Table 3, we report the correlations among the main independent variables and pre-measures (i.e., Evaluation pressure, MKT, social norms and Fall TRU Math). Only MKT scores had significant positive correlations with some Fall TRU Math scores, which suggests potential multicollinearity problem. Accordingly, we attended to the changes in the standard errors especially those of MKT and pre-measures in the following analysis, but there was no sign of a serious multicollinearity issue.

Table 3. Correlation among the main teacher level variables

	(1)	(2)	(3)	(4)	(5)	(6)
(1) Evaluation Pressure						
(2) MKT	0.173					
(3) Social norms	-0.023	-0.133				
(4) Dimension 1 in Fall	-0.153	0.185	-0.160			
(5) Dimension 2 in Fall	-0.173	0.184	-0.163	0.841***		
(6) Dimension 3 in Fall	-0.064	0.245*	-0.102	0.692***	0.738***	
(7) Dimension 4 in Fall	-0.119	0.378***	-0.123	0.697***	0.736***	0.902***

Note. All ratings were aggregated at the teacher level. *$p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

With regard to the first research question, Table 4 presents the association between teachers' perceived pressure associated with teacher evaluation and changes in teachers' ambitious mathematics instruction. The pressure an ECT perceived had a significant and negative association with changes in their ambitious mathematics instruction measured by dimensions 2 (cognitive demand) and 4 (uses of assessment), and combined scores. However, across all dimension scores, teachers' perceived pressure had negative associations with the outcome.

Putting such results into the language of the rubric, ECTs who perceived high levels of pressure related to teacher evaluation with regard to their instruction were less likely to teach

cognitively demanding tasks (dimension 2); and these ECTs were less likely to monitor student ideas or use them in the class (dimension 4). The magnitude of the association was stronger for dimension 2 scores; a one-unit increase in teachers' perceived pressure of teacher evaluation was associated with a 0.169-point lower TRU Math score controlling for various control variables as well as pre-measures and district fixed effects. Given the considerable difference in 1 point based on the rubric (See Appendix A for details), this value is not negligible. This is equivalent to 30.95% of a standard deviation. The magnitude of the associations between evaluation influence and dimension 4 and the combined scores are similar but slightly less than those between teacher evaluation pressure and the dimension 2 ratings. The models in Tables 5 and 6 included MKT and social norms respectively and the findings stayed almost the same. Neither MKT nor social norms had a significant association with the outcomes in these models. Our final models are in Table 7, where we included both MKT and social norms to estimate the association between ECTs' perceived evaluation pressure and mathematics instruction. The association becomes stronger and dimension 1 (mathematics) ratings also showed a significant association with the pressure variable.

Table 4. Estimated Effects of Teachers' Perceived Pressure Associated with Teacher Evaluation on Teachers' Enactment of Mathematics Instruction

	Model (1) Dimension1: The mathematics	Model (2) Dimension2: Cognitive demand	Model (3) Dimension3: Agency, authority, and identity	Model (4) Dimension4: Uses of assessment	Model (5) Four Dimensions Combined
Mean score in fall 2015	0.298***	0.259**	0.181	0.241**	0.301***
	(0.090)	(0.085)	(0.093)	(0.091)	(0.086)
Evaluation pressure	-0.122	-0.169**	-0.079	-0.127*	-0.114*
	(0.065)	(0.061)	(0.051)	(0.055)	(0.050)
Missing flag_evaluation pressure	-1.230*	-1.794***	-0.965*	-1.799***	-1.428***
	(0.482)	(0.451)	(0.378)	(0.407)	(0.369)
Video observation	-0.273**	-0.275**	-0.024	-0.131	-0.182*
	(0.095)	(0.089)	(0.075)	(0.081)	(0.073)
Teaching a tested grade	0.075	0.068	-0.025	-0.007	0.041
	(0.077)	(0.073)	(0.061)	(0.065)	(0.059)
Number of students in class	-0.001	9.09e-05	0.022**	0.021*	0.007
	(0.010)	(0.009)	(0.008)	(0.009)	(0.008)
Professional development	-0.041	-0.041*	-0.008	-0.012	-0.027
	(0.022)	(0.021)	(0.017)	(0.019)	(0.017)
Hold a Master's degree	-0.015	-0.069	-0.066	-0.047	-0.037
	(0.086)	(0.081)	(0.068)	(0.073)	(0.066)
Years of teaching experience	0.064*	-0.008	-0.008	-0.0005	0.016
	(0.028)	(0.026)	(0.022)	(0.023)	(0.021)
Constant	1.659***	1.989***	0.746**	0.943**	1.288***
	(0.352)	(0.321)	(0.262)	(0.290)	(0.267)
Number of episodes	1,117	1,109	1,098	1,101	1,141
Number of teachers	96	96	96	96	96

Note. All models included district fixed effects and standard errors in parentheses. Missing flags for professional development, whether a teacher held a Master's degree, and total years of teaching experience were also included all models. *$p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Table 5. Estimated Effects of Teachers' Perceived Pressure Associated with Teacher Evaluation and MKT on Teachers' Enactment of Mathematics Instruction

	Model (1) Dimension1: The mathematics	Model (2) Dimension2: Cognitive demand	Model (3) Dimension3: Agency, authority, and identity	Model (4) Dimension4: Uses of assessment	Model (5) Four Dimensions Combined
Mean score in fall 2015	0.317***	0.254**	0.171	0.264**	0.301***
	(0.093)	(0.089)	(0.093)	(0.093)	(0.089)
Evaluation pressure	-0.117	-0.170**	-0.089	-0.130*	-0.117*
	(0.066)	(0.062)	(0.051)	(0.055)	(0.050)
Missing flag_evaluation pressure	-1.184*	-1.758***	-0.956*	-1.732***	-1.380***
	(0.481)	(0.453)	(0.377)	(0.400)	(0.366)
MKT	-0.050	-0.030	-0.004	-0.064	-0.041
	(0.039)	(0.036)	(0.030)	(0.033)	(0.030)
Video observation	-0.287**	-0.287**	-0.036	-0.157	-0.200**
	(0.096)	(0.091)	(0.076)	(0.080)	(0.073)
Teaching a tested grade	0.078	0.070	-0.019	0.003	0.047
	(0.077)	(0.073)	(0.061)	(0.064)	(0.059)
Number of students in class	-0.001	0.0004	0.022**	0.021*	0.007
	(0.010)	(0.009)	(0.008)	(0.008)	(0.008)
Professional development	-0.042	-0.041*	-0.009	-0.015	-0.028
	(0.022)	(0.021)	(0.017)	(0.018)	(0.017)
Hold a Master's degree	-0.008	-0.064	-0.064	-0.036	-0.030
	(0.086)	(0.082)	(0.068)	(0.072)	(0.066)
Years of teaching experience	0.0578*	-0.011	-0.006	-0.006	0.013
	(0.028)	(0.026)	(0.022)	(0.023)	(0.021)
Constant	1.640***	2.013***	0.783**	0.958***	1.318***
	(0.356)	(0.326)	(0.261)	(0.287)	(0.268)
Number of episodes	1,106	1,100	1,089	1,092	1,130
Number of teachers	95	95	95	95	95

Note. All models included district fixed effects and standard errors in parentheses. Missing flags for professional development, whether a teacher held a Master's degree, and total years of teaching experience were also included all models. *$p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Table 6. Estimated Effects of Teachers' Perceived Pressure Associated with Teacher Evaluation and social norms on Teachers' Enactment of Mathematics Instruction

	Model (1) Dimension1: The mathematics	Model (2) Dimension2: Cognitive demand	Model (3) Dimension3: Agency, authority, and identity	Model (4) Dimension4: Uses of assessment	Model (5) Four Dimensions Combined
Mean score in fall 2015	0.302***	0.277**	0.182*	0.248**	0.315***
	(0.087)	(0.085)	(0.093)	(0.091)	(0.086)
Evaluation pressure	-0.139*	-0.177**	-0.081	-0.131*	-0.121*
	(0.063)	(0.060)	(0.051)	(0.055)	(0.049)
Missing flag_evaluation pressure	-1.122*	-1.748***	-0.942*	-1.785***	-1.387***
	(0.467)	(0.446)	(0.380)	(0.408)	(0.364)
Social norms	0.005	0.008	-0.003	0.005	0.006
	(0.012)	(0.012)	(0.010)	(0.011)	(0.010)
Missing flag_social norms	-0.238**	-0.139	-0.033	-0.051	-0.113
	(0.090)	(0.087)	(0.074)	(0.079)	(0.071)
Video observation	-0.295**	-0.288**	-0.027	-0.135	-0.192**
	(0.092)	(0.088)	(0.075)	(0.081)	(0.072)
Teaching a tested grade	0.086	0.078	-0.026	-0.003	0.048
	(0.075)	(0.072)	(0.061)	(0.066)	(0.059)
Number of students in class	0.0001	0.0009	0.022**	0.021*	0.008
	(0.010)	(0.009)	(0.008)	(0.009)	(0.008)
Professional development	-0.043*	-0.043*	-0.007	-0.014	-0.029
	(0.021)	(0.021)	(0.017)	(0.019)	(0.017)
Hold a Master's degree	-0.066	-0.101	-0.071	-0.060	-0.063
	(0.085)	(0.082)	(0.070)	(0.075)	(0.067)
Years of teaching experience	0.080**	0.0003	-0.005	0.003	0.024
	(0.027)	(0.026)	(0.022)	(0.024)	(0.021)
Constant	1.611***	1.902***	0.769**	0.898**	1.229***
	(0.354)	(0.330)	(0.272)	(0.302)	(0.275)
Number of episodes	1,117	1,109	1,098	1,101	1,141
Number of teachers	96	96	96	96	96

Note. All models included district fixed effects and standard errors in parentheses. Missing flags for professional development, whether a teacher held a Master's degree, and total years of teaching experience were also included all models. *$p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Table 7. Estimated Effects of Teachers' Perceived Pressure Associated with Teacher Evaluation, MKT, and social norms on Teachers' Enactment of Mathematics Instruction

	Model (1) Dimension1: The mathematics	Model (2) Dimension2: Cognitive demand	Model (3) Dimension3: Agency, authority, and identity	Model (4) Dimension4: Uses of assessment	Model (5) Four Dimensions Combined
Mean score in fall 2015	0.299***	0.259**	0.175	0.265**	0.302***
	(0.090)	(0.088)	(0.092)	(0.092)	(0.087)
Evaluation pressure	-0.145*	-0.186**	-0.097	-0.139*	-0.132**
	(0.064)	(0.062)	(0.052)	(0.055)	(0.050)
Missing flag_evaluation	-1.048*	-1.693***	-0.909*	-1.695***	-1.314***
pressure	(0.466)	(0.448)	(0.378)	(0.400)	(0.361)
MKT	-0.050	-0.029	-0.005	-0.063	-0.040
	(0.037)	(0.036)	(0.030)	(0.033)	(0.029)
Social norms	0.004	0.007	-0.004	0.003	0.004
	(0.012)	(0.012)	(0.010)	(0.010)	(0.009)
Missing flag_social norms	-0.261**	-0.162	-0.075	-0.085	-0.146*
	(0.095)	(0.091)	(0.077)	(0.081)	(0.073)
Video observation	-0.320***	-0.307***	-0.046	-0.167*	-0.219**
	(0.093)	(0.090)	(0.076)	(0.081)	(0.072)
Teaching a tested grade	0.092	0.081	-0.017	0.008	0.056
	(0.074)	(0.072)	(0.061)	(0.064)	(0.058)
Number of students in class	0.0004	0.001	0.022**	0.021*	0.008
	(0.010)	(0.009)	(0.008)	(0.008)	(0.007)
Professional development	-0.044*	-0.044*	-0.008	-0.016	-0.030
	(0.021)	(0.020)	(0.017)	(0.018)	(0.016)
Hold a Master's degree	-0.061	-0.100	-0.075	-0.054	-0.061
	(0.085)	(0.082)	(0.069)	(0.073)	(0.066)
Years of teaching experience	0.078**	0.001	0.001	0.001	0.024
	(0.028)	(0.027)	(0.023)	(0.024)	(0.022)
Constant	1.665***	1.965***	0.812**	0.942**	1.297***
	(0.358)	(0.335)	(0.270)	(0.297)	(0.274)
Number of episodes	1,106	1,100	1,089	1,092	1,130
Number of teachers	95	95	95	95	95

Note. All models included district fixed effects and standard errors in parentheses. Missing flags for professional development, whether a teacher held a Master's degree, and total years of teaching experience were also included all models. *$p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Perceived Pressure Related to Teacher Evaluation, MKT, and Social Norms

The second and third research questions asked whether the association between evaluation pressure and ambitious instructions is shaped by resources at the individual teacher level (i.e., MKT) and/or at the school level (i.e., social norms). Table 8 summarizes the results for the second research question. The interaction term between teachers' MKT and evaluation pressure had a negative and significant association with changes in teachers' TRU Math scores for dimensions 3 and 4, as well as combined ratings. It should be noted that in these models, MKT and evaluation pressure variables are grand-mean centered in order to examine the association of these variables for the teachers whose MKT and evaluation pressure are at the mean. That is, for teachers who perceived an average level of evaluation pressure, a one-unit increase in their MKT score is associated with a substantial decrease in their enactment of ambitious mathematics instruction in terms of dimensions 3 and 4, and the combined measure. The coefficients for dimensions 1 and 2 were also negative, but not statistically significant.

Table 8. Estimated Effects of Teachers' Perceived Pressure Associated with Teacher Evaluation on Teachers' Enactment of Mathematics Instruction: Heterogeneous Effects Based on Teachers' MKT

	Model (1) Dimension1: The mathematics	Model (2) Dimension2: Cognitive demand	Model (3) Dimension3: Agency, authority, and identity	Model (4) Dimension4: Uses of assessment	Model (5) Four Dimensions Combined
Mean score in fall 2015	0.290**	0.254**	0.201*	0.294***	0.305***
	(0.090)	(0.087)	(0.089)	(0.087)	(0.084)
Evaluation pressure	-0.178**	-0.213**	-0.147**	-0.208***	-0.179***
	(0.068)	(0.066)	(0.053)	(0.055)	(0.051)
Missing flag_evaluation pressure	-1.004*	-1.658***	-0.853*	-1.612***	-1.257***
	(0.464)	(0.446)	(0.364)	(0.378)	(0.350)
MKT	-0.050	-0.029	-0.006	-0.065*	-0.041
	(0.037)	(0.036)	(0.029)	(0.031)	(0.028)
Social norms	0.002	0.006	-0.006	4.80e-05	0.002
	(0.012)	(0.012)	(0.009)	(0.010)	(0.009)
Missing flag_social norms	-0.270**	-0.169	-0.091	-0.105	-0.158*
	(0.095)	(0.091)	(0.074)	(0.076)	(0.071)
Video observation	-0.334***	-0.319***	-0.071	-0.197**	-0.239***
	(0.093)	(0.090)	(0.073)	(0.076)	(0.070)
Teaching a tested grade	0.091	0.080	-0.014	0.013	0.057
	(0.074)	(0.072)	(0.059)	(0.060)	(0.056)
Number of students in class	0.0008	0.002	0.022**	0.021**	0.008
	(0.009)	(0.009)	(0.008)	(0.008)	(0.007)
Professional development	-0.043*	-0.043*	-0.007	-0.014	-0.029
	(0.021)	(0.020)	(0.017)	(0.017)	(0.016)
Hold a Master's degree	-0.060	-0.099	-0.073	-0.052	-0.060
	(0.084)	(0.081)	(0.067)	(0.069)	(0.064)
Years of teaching experience	0.076**	-0.001	-0.002	-0.003	0.021
	(0.028)	(0.027)	(0.022)	(0.023)	(0.021)
MKT*Evaluation pressure	-0.109	-0.092	-0.181**	-0.250***	-0.162**
	(0.083)	(0.080)	(0.066)	(0.068)	(0.063)
Constant	1.777***	2.051***	0.945***	1.123***	1.436***
	(0.365)	(0.342)	(0.264)	(0.283)	(0.271)
Number of episodes	1,106	1,100	1,089	1,092	1,130
Number of teachers	95	95	95	95	95

Note. All models included district fixed effects and standard errors in parentheses. Missing flags for professional development, whether a teacher held a Master's degree, and total years of teaching experience were also included all models. MKT and evaluation pressure are grand-mean centered to ease interpretation. *$p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

As noted earlier, however, this result can be interpreted in a different way; it is plausible that teachers' perceptions of teacher evaluation interfered with the translation of MKT into instruction. That is, this result may explain why there was no significant association between MKT and TRU Math in the models. We argue that teacher evaluation pressure may weaken the association between the two as MKT might have led teachers to teach differently based on their level of perceived pressure associated with teacher evaluation.

Lastly, in all models, social norms at a given school did not have any influence on the association between evaluation pressure and changes in teachers' ambitious mathematics instruction. In contrast to our hypothesis, this finding suggests that social norms do not buffer or accelerate the negative association between teacher evaluation pressure and ECTs' instructional practices. This result stayed the same when we excluded evaluation pressure from the model (the results are available upon request).

Robustness Check

To test the robustness of our findings, we ran models with different specifications. See Appendix C for detailed results. First, we ran models without imputed variables. As noted earlier, we imputed missing values for evaluation pressure, social norms, professional development, whether a teacher held a master's degree, and years of teaching experience. In order to understand whether such decisions changed our findings, we employed list-wise deletion where we dropped all of the cases that had one or more missing values. Although the sample size shrank, the association became stronger for all dimension scores. For these models, dimension 3 scores also had a significant negative association with perceived evaluation pressure. This pattern stayed the same when we dropped all district fixed effects.

Another main decision that we made was about the level of analysis. In this regard, we ran the main models with teacher-level aggregated variables with and without school fixed effects. Across all models, the findings are consistent with the previous findings; in many models, the association between the two main variables was stronger than the one from the main models.

On the other hand, it is possible that there are omitted variables that might have introduced bias in our findings. According to Robustness Indices (Frank, 2000; Frank, Maroulis, Duong, & Kelcey, 2013), however, to invalidate the inference in this study between teachers' ambitious mathematics instruction and perceived pressure related to teacher evaluation, 34% of the sample for dimension 2 (cognitive demand), and 24% of the sample for the combined measure would need to be replaced with cases that show no association between the two. These values are at a similar level of the median of other studies reported by Frank et al. (2013). In terms of potential heterogeneous effects of teachers' perceived pressure of teacher evaluation based on MKT, 46% of the sample for dimension 4 (uses of assessment), and 23% of the sample for the combined measure would need to be replaced with cases for which the effects of the two variables (i.e., pressure of evaluation and MKT) on the outcome were purely additive. In conclusion, we believe that the association between teachers' perceived evaluation pressure and their mathematics instruction is rather robust to different specifications and plausible omitted variables.

Discussion

Over the last decade, both the CCSS and teacher evaluation policies have gained significant attention as a means to enhance teaching quality. Therefore, understanding whether and how these two forces interact to impact teaching and learning is essential (Coburn et al.,

2016). Our findings indicate that current teacher evaluation policies might conflict with the ambitious instruction movement that CCSS and many teacher preparation programs promote. This is consistent with Polikoff and Porter's (2014) finding of a weak correlation between teachers' instructional alignment with state standards and their contribution to student test scores and Welch and colleagues' (2016) finding of little overlap between what teacher evaluation rubrics measured and what CCSS emphasized.

This finding does not seem to be a coincidence; as we showed with our document analysis, what the current observation rubrics measure may not be well-aligned with the elements of ambitious instruction. In other words, there was not enough affordance for enacting ambitious instruction under the current teacher evaluation policies and, thus, teachers may have made a rational decision, in favor of positive feedback from teacher evaluation and job security, to move away from enacting ambitious mathematics instruction.

This finding is consistent with some teachers' responses to high-stakes school accountability policies; with limited time and other resources, teachers strategically spent their time and resources to meet the requirements of *No Child Left Behind* and similar policies (Booher-Jennings, 2005; Jacob, 2005; Reback et al., 2014; White & Rosenbaum, 2008). The current study confirms that the same issue might pertain to teacher evaluation settings. Moreover, given that ECTs working in the 2015-16 school year are more likely to experience mathematics methods courses focused on ambitious mathematics instruction in their preparation programs, it is alarming that the effects of these courses are likely diminished when graduates begin working as full-time teachers under the current teacher evaluation policies.

The results regarding the potential moderating effects of teachers' MKT shed light on a more nuanced aspect of the potential influence of teacher evaluation on teachers' instructional

practices. First, although teachers' MKT has been argued to be critical for high-quality mathematics teaching (Hill et al., 2005; Hill et al., 2007), with strong pressure from teacher evaluation, the association between MKT and ambitious mathematics teaching in this study was weak. From a different perspective, this can also be interpreted as high MKT exacerbating the potential negative impact of teacher evaluation policies on ambitious mathematics instruction. Under both interpretations, ECTs seem to use their teaching expertise to make their instruction more aligned with teacher evaluation rather than ambitious mathematics teaching. Interestingly, Steinberg and Sartain (2015) argued that "higher human capital teachers are likely more able to incorporate principal feedback and assessment into their instructional practice" (p. 566). Contrary to this, we found that incorporating principal feedback and assessment into practice may not be well-aligned with ambitious mathematics instruction and this was more salient with teachers who had greater levels of human capital.

On the other hand, however, this finding means that for teachers whose MKT is relatively low, teacher evaluation pressure may encourage them to grow. This finding suggests a potential tension between two possible goals of teacher evaluation; improving struggling teachers at the bottom versus high-performers at the top. Although this part of the analysis was exploratory due to the small number of teachers in our sample, this issue deserves further analysis especially in terms of ways to maximize the effects of teacher evaluation for teachers at different spectrum of effectiveness.

There are a few limitations of the current study that can be addressed in future studies. First of all, we acknowledge that teachers' perceived pressure associated with teacher evaluation may not be perfectly consistent with the actual pressure emanating from teacher evaluation policy. For example, we do not have data on ECTs' actual teacher evaluation ratings to gauge

how important it would be for them to achieve high ratings and/or to keep their current positions. Moreover, we do not have detailed information about how much emphasis each district placed on teacher evaluation, which is often unclear in formal policy documents. That is, we do not know how much actual pressure each teacher experienced. However, building on a long history of literature on policy implementation (McLaughlin, 1987), we argue that teachers' perceptions of policy are a valid measure of policy pressure and that they directly affect teachers' behavioral responses. Formal policies or potential sanctions certainly affect teachers' behavior, but such external forces go through teachers' sensemaking processes (Coburn, 2001) to make changes in their behavior. In the current study, we directly measured how much pressure the policy had on teachers from their perspective, rather than presuming low ratings or a specific aspect of policy would mean a stronger influence of policy. On the other hand, in order to understand how the policy seems to affect teachers' actual instructional practice, we drew on observation data, rather than relying on teachers' self-reports of how they changed their behavior due to teacher evaluation.

Second, we acknowledge that our sample is self-selected; ECTs who were willing to open their classroom to researchers and to complete multiple surveys and interviews chose to participate in the project. Similarly, we do not have any data to show how representative the social networks members of ECTs who completed the surveys were. We admit that this self-selection nature of the current study puts limits on generalizing the findings. Due to the lack of data, it was not possible to compare the characteristics of respondents and non-respondents, but we compared our ECT sample with a nationally representative career teacher sample to understand this issue better. Most observational studies, especially studies that utilize classroom

observation data, are not free from this potential bias, which warrants efforts to develop less intrusive ways of collecting authentic data on teaching practices.

Third, other school-level factors might affect this association between the teacher evaluation pressure and teacher instruction. Specifically, school leadership can shape the influence of teacher evaluation policies. For example, school administrators who enact strong instructional leadership may strengthen or weaken the influence of teacher evaluation policies, given the critical role of principals in the teacher evaluation process (Delvaux et al., 2013). Studying these factors might add new insights for studies of the influence of teacher evaluation policies.

Implications and Conclusion

This study provides empirical evidence that shows misalignment between current teacher evaluation policies and ambitious instruction and potential heterogeneity in the influence of teacher evaluation. The most current reauthorization of the *Elementary and Secondary Education Act*, the 2015 *Every Student Succeeds Act*, granted states a high level of latitude in evaluating educators while they strive to ensure that all students have access to high quality teaching (Egalite, Fusarelli, & Fusarelli, 2017). An important policy question to consider at this juncture is how states and districts can design evaluation systems that serve summative and formative purposes while promoting ambitious instruction among all teachers. This question certainly deserves further analysis and discussion beyond this study, but we suggest two things to consider in developing such a coherent system based on our findings. First, simply providing teachers with more training to enhance knowledge for teaching would not be enough. Without a support system that helps them appropriate their knowledge, which is an essential part of their expertise in teaching, such knowledge will not be translated into high-quality teaching. Second, at least for

ECTs, teacher evaluation seems to be a powerful tool to shape teachers' instruction, although only a few teachers have been subject to rewards or sanctions (Kraft, & Gilmour, 2017). Thus, it is essential to understand the direction in which teacher evaluation leads teachers, which is oftentimes ignored in debates about teacher evaluation policies.

Endnotes

1. Compared to the nationally representative sample of early career teachers in the National Teacher and Principal Survey, the teachers in our sample were more likely to be female, (χ^2 (1, N = 2919) = 6.37, p < .05) more likely to have a general teaching certification (χ^2 (1, N = 2919) =41.74, p < .001), less likely to be Hispanic (χ^2 (1, N = 2919) =9.35, p < .01) and more likely have majored in education(χ^2 (1, N = 2919) =38.71, p < .001), but there was no statistically significant difference in whether they held master's degree (Frank et al., forthcoming). We acknowledge that this discrepancy might limit our ability to generalize the findings from this study.
2. For example, assume that ECT A nominated Teacher B as a network member in the fall, but A did not nominate B as a network member in the spring. In that case, we did not include Teacher B's data in the analysis unless another ECT at the same school nominated Teacher B in the spring. On the other hand, if an ECT nominated Teacher C only in the spring, not in the fall, and none of the other ECTs at the same school nominated Teacher C in the fall, we excluded Teacher C from the analysis.
3. When two of the co-authors independently coded alignment between teacher evaluation and the CCSS based on the rubric developed by Welch et al. (2016) in the first round, the percentage of agreement was 77.78%. After having a discussion, the coders reached agreement for all dimensions of the rubric.

References

Ball, D. L. & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In G. Sykes & L. Daling-Hammond (Eds,), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3-32). San Francisco: Jossey Bass.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389-407.

Bell, C., Qi, Y., Croft, A., Leusner, D., McCaffrey, D., Gitomer, D., & Pianta, R. (2014). Improving observational score quality: Challenges in observer thinking. In K. Kerr, R. Pianta, & T. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 50–97). San Francisco, CA: Jossey-Bass.

Bieda, K., Salloum, S., Hu, S., Sweeny, S., Torphy, K., & Lane, J. (in press). Issues with, and insights from, capturing the quality of mathematics teaching at scale. *Journal of Classroom Interaction*.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review, 48*, 16-29.

Blazar, D., & Pollard, C. (2017). Does test preparation mean low-quality instruction? *Educational Researcher*, *46*(8), 420-433.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, *42*(2), 231–268.

Bridwell-Mitchell, E. N., & Sherer, D. G. (2017). Institutional complexity and policy implementation: How underlying logics drive teacher interpretations of reform. *Educational Evaluation and Policy Analysis*, *39*(2), 223-247.

Brown, C. A., Stein, M. K., & Forman, E. A. (1996). Assisting teachers and students to reform the mathematics classroom. *Educational Studies in Mathematics*, *31*(1-2), 63-93.

Cobb, P., Boufi, A., McClain, K., & Whitenack, J. W. (1997). Reflective discourse and collective reflection. *Journal for Research in Mathematics Education, 28*, 258-277.

Cobb, P., & Smith, T. (2008). District development as a means of improving mathematics teaching and learning at scale. In K. Krainer, & T. Wood (Eds.), *International handbook of mathematics teacher education: Participants in mathematics teacher education* (Vol. 3, pp. 231–254). Rotterdam: Sense Publishers.

Coburn, C. E. (2001). Collective sensemaking about reading: How teachers mediate reading policy in their professional communities. *Educational Evaluation and Policy Analysis*, *23*(2), 145–170.

Coburn, C. E., Hill, H. C., & Spillane, J. P. (2016). Alignment and accountability in policy design and implementation: The Common Core State Standards and implementation research. *Educational Researcher*, *45*(4), 243-251.

Coburn, C. E., & Russell, J. L. (2008). District policy and teachers' social networks. *Educational Evaluation and Policy Analysis*, *30*(3), 203-235.

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

Consortium on Chicago School Research. University of Chicago. (2017]. 5Essentials® Survey. Chicago, IL: University of Chicago Consortium on School Research

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*(4), 724-750.

Danielson, C. (2013). *The Framework for Teaching evaluation instrument.* Princeton, NJ: The Danielson Group.

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

Delvaux, E., Vanhoof, J., Tuytens, M., Vekeman, E., Devos, G., & Van Petegem, P. (2013). How may teacher evaluation have an impact on professional development? A multilevel analysis. *Teaching & Teacher Education*, *36*, 1–11.

Donaldson, M. (2012). *Teachers' perspectives on teacher evaluation reform*. Washington, DC: Center for American Progress.

Donaldson, M. L., Woulfin, S., LeChasseur, K., & Cobb, C. D. (2016). The structure and substance of teachers' opportunities to learn about teacher evaluation reform: promise or pitfall for equity? *Equity & Excellence in Education*, *49*(2), 183-201.

Egalite, A. J., Fusarelli, L. D., & Fusarelli, B. C. (2017). Will decentralization affect educational inequity? The Every Student Succeeds Act. *Educational Administration Quarterly*, *53*(5), 757-781.

Feiman-Nemser, S., & Buchman, M. (1985). Pitfalls of experience in teacher preparation. *Teachers College Record*, *87*(1), 53-65.

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, *29*(2), 147-194.

Frank, K.*, Kim, J.*, Salloum, S., Bieda, K., & Youngs, P. (accepted). From interpretation to instructional practice: A network study of early career teachers' sensemaking in the era of accountability pressures and Common Core State Standards. *American Educational Research Journal*. *Co-equal first authorship

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, *35*(4), 437-460.

Frank, K. A., Zhao, Y., Penuel, W. R., Ellefson, N. C., & Porter, S. (2011). Focus, fiddle and friends: Sources of knowledge to perform the complex task of teaching. *Sociology of Education*, *84*(2), 137-156.

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

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

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

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*(1), 5-28.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., & Robyn, A. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. Santa Monica, CA: Rand Corporation.

Hill, H. (2011). *Mathematical Quality of Instruction*. Retrieved from https://cepr.harvard.edu/ncte-conference-2011

Hill, H. C., Ball, D. L., Blunk, M., Goffney, I. M., & Rowan, B. (2007). Validating the ecological assumption: The relationship of measure scores to classroom teaching and student learning. *Measurement*, *5*(2-3), 107-118.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Hill, H.C., Schilling, S.G., & Ball, D.L. (2004) Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal 105*(1), 11-30.

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics*, *89*(5), 761–796.

Jennings, N. E., & Spillane, J. P. (1996). State reform and local capacity: Encouraging ambitious instruction for all and local decision-making. *Journal of Education Policy, 11*(4), 465-482.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform Chicago's REACH students. *Educational Researcher*, *44*(2), 105–116.

Jiang, J. & Sporte, S. (2016). Teacher evaluation in Chicago: Differences in observation and value-added scores by teacher, student, and school characteristics. Chicago, IL: The University of Chicago

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learning mathematics.* Washington, DC: The National Academies Press.

Kim, J., Sun, M., & Youngs, P. (2019). Developing the "will": The relationship

between teachers' perceived policy legitimacy and instructional improvement. *Teachers College Record*, 121(3), 1-44.

Koedel, C., Li, J., Springer, M. G., & Tan, L. (2019). Teacher performance ratings and professional improvement. *Journal of Research on Educational Effectiveness*, *12*(1), 90-115.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*(5), 234-249.

Lampert, M., Beasley, H., Ghousseini, H., Kazemi, E., & Franke, M. (2010). Using designed instructional activities to enable novices to manage ambitious mathematics teaching. In M. K. Stein & L. Kucan (Eds.), *Instructional explanations in the disciplines* (pp. 129-141). New York: Springer.

Lampert, M., Franke, M. L., Kazemi, E., Ghousseini, H., Turrou, A. C., Beasley, H., Cunard, A., & Crowe, K. (2013). Keeping it complex: Using rehearsals to support novice teacher learning of ambitious teaching. *Journal of Teacher Education*, *64*(3), 226-243.

Lave, J. (1996). Teaching, as learning, in practice. *Mind, Culture, and Activity, 3*(3), 149-164.

Lee, E., Brown, M. N., Luft, J. A., & Roehrig, G. H. (2007). Assessing beginning secondary science teachers' PCK: Pilot year results. *School Science and Mathematics*, *107*(2), 52-60.

Leko, M., & Brownell, M. T. (2011). Understanding the various influences on special education preservice teachers' appropriation of conceptual and practical tools for teaching reading. *Exceptional Children, 77*(2), 229-251.

Marsh, J. A., Bush-Mecenas, S., Strunk, K. O., Lincove, J. A., & Huguet, A. (2017). Evaluating teachers in the Big Easy: How organizational context shapes policy responses in New

Orleans. *Educational Evaluation and Policy Analysis*, *39*(4), 539-570.

McClain, K. (2002). Teacher's and students' understanding: The role of tool use in communication. *Journal of the Learning Sciences, 11,* 217-249.

McLaughlin, M. W. (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis*, *9*(2), 171–178.

Milanowski, A. T., & Heneman, H. G. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education*, *15*(3), 193–212.

Penuel, W. R., Sun, M., Frank, K. A., & Gallagher, H. A. (2012). Using social network analysis to study how collegial interactions can augment teacher learning from external professional development. *American Journal of Education*, *119*(1), 103-136.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*, *36*(4), 399-416.

Ramirez, A., Clouse, W., & Davis, K. W. (2014). Teacher evaluation in Colorado: How policy frustrates practice. *Management in Education*, *28*(2), 44-51.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy*, *6*(3), 207–241.

Reinhorn, S. K., Johnson, S. M., & Simon, N. S. (2017). Investing in development: Six high-performing, high-poverty schools implement the Massachusetts teacher evaluation policy. *Educational Evaluation and Policy Analysis*, *39*(3), 383-406.

Rogoff, B. (1994). Developing understanding of the idea of communities of learners. *Mind, Culture, and Activity, 1*(4), 209-229.

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM*, *45*(4), 607-621.

Schoenfeld, A. H., Floden, R. E., & the Algebra Teaching Study and Mathematics Assessment Project. (2014). *The TRU Math scoring rubric*. Berkeley, CA & E. Lansing, MI: Graduate School of Education, University of California, Berkeley & College of Education, Michigan State University. Retrieved from http://ats.berkeley.edu/tools.html

Smith, T. M., Booker, L., Hochberg, E., & Desimone, L. M. (2018). Do organizational supports for mathematics instruction improve the quality of beginning teachers' instruction? *Teachers College Record*, *120*(7) 1-46.

Spillane, J. P., & Jennings, N. E. (1997). Aligned instructional policy and ambitious pedagogy: Exploring instructional reform from the classroom perspective. *Teachers College Record, 98*(3), 449-81.

Spillane, J., Reiser, B., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of Educational Research, 72*(3), 387–431.

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. *Santa Monica, CA: RAND Corporation. Retrieved from www. rand. org/pubs/research_reports/RR2242.html*

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 340-359.

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

Stroupe, D. (2016). Beginning teachers' use of resources to enact and learn from ambitious instruction. *Cognition and Instruction*, *34*(1), 51-77.

Sun, M., Frank, K. A., Penuel, W. R., & Kim, C. M. (2013). How external institutions penetrate schools through formal and informal leaders. *Educational Administration Quarterly*, *49*(4), 610-644.

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

Thames, M. H., & Ball, D. L. (2010). What math knowledge does teaching require? *Teaching Children Mathematics, 17*(4), 220-229.

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

Tuytens, M., & Devos, G. (2009). Teachers' perception of the new teacher evaluation policy: A validity study of the policy characteristics scale. *Teaching and Teacher Education*, *25*(6), 924–930.

Tuytens, M., & Devos, G. (2010). The influence of school leadership on teachers' perception of teacher evaluation policy. *Educational Studies*, *36*(5), 521–536.

Welch, M., Davis, S., Isaia, R., Johnston, W., Stein, L., Jenuwine, H., & Macdonald, K. (2016). Aligning evaluation: How much do teacher evaluation rubrics emphasize common core instruction? Derived from

https://www.air.org/sites/default/files/downloads/report/Teacher-Evaluation-Common-Core-Alignment-October-2016.pdf

White, K. & Rosenbaum, J. (2008). Inside the black box of accountability: How high-stakes accountability alters school culture and the classification and treatment of students and teachers. In A. R. Sadovnik, J. A. O'Day, G. W. Bohrnstedt, & K. M. Borman (Eds.), *No Child Left Behind and the reduction of the achievement gap: Sociological perspectives on federal education policy* (pp. 97–116). New York: Routledge.

Zeichner, K. M., & Tabachnick, B. R. (1981). Are the effects of university teacher education 'washed out' by school experience? *Journal of Teacher Education*, *32*(3), 7-11.