5-28-2017

U.S. Principals' Interpretation and Implementation of Teacher Evaluation Policies

David B. Reid
Michigan State University, reiddav1@msu.edu

Follow this and additional works at: http://nsuworks.nova.edu/tqr

Part of the Educational Assessment, Evaluation, and Research Commons, and the Educational Leadership Commons

U.S. Principals' Interpretation and Implementation of Teacher Evaluation Policies

Abstract

In the United States policymakers, states, and researchers are increasingly reliant on teacher evaluations as a means for identifying high-quality teachers. School principals are the primary school-based actors responsible for implementing teacher evaluation policies at the school level and must make sense of these policies at an ever-increasing pace. These sensemaking processes have great implications for how teacher evaluation policies play out in practice. In this paper I ask (a) what factors influence principals' sensemaking of changing teacher evaluation policies and (b) how these factors influence both decision-making by principals, as well as the ways the policies are implemented. I use an exploratory case study approach, drawing on interviews and district specific documents from six public school principals in the U.S. state of Michigan. Findings suggest that, because teacher evaluation policies were tied to the employment of their teachers, principals made sense of and implemented these policies in very specific ways. Implications for policy and practice are discussed.

U.S. Principals' Interpretation and Implementation of Teacher Evaluation Policies

David B. Reid

Michigan State University, East Lansing, Michigan, USA

In the United States policymakers, states, and researchers are increasingly reliant on teacher evaluations as a means for identifying high-quality teachers. School principals are the primary school-based actors responsible for implementing teacher evaluation policies at the school level and must make sense of these policies at an ever-increasing pace. These sensemaking processes have great implications for how teacher evaluation policies play out in practice. In this paper I ask (a) what factors influence principals' sensemaking of changing teacher evaluation policies and (b) how these factors influence both decision-making by principals, as well as the ways the policies are implemented. I use an exploratory case study approach, drawing on interviews and district specific documents from six public school principals in the U.S. state of Michigan. Findings suggest that, because teacher evaluation policies were tied to the employment of their teachers, principals made sense of and implemented these policies in very specific ways. Implications for policy and practice are discussed. Keywords: Teacher Evaluation, Policy Implementation, Principal, Sensemaking

Most researchers and policymakers agree one step towards overall school improvement and the long-term success of students is access to quality teaching (Chetty, Friedman, & Rockoff, 2014; Darling-Hammond, 2000). Despite the consensus on the importance of teacher quality, current teacher evaluation systems in the U.S. have fallen short in identifying high-quality teachers. For example, in 2014, 97 percent of teachers in the U.S. state of Michigan were rated as effective or highly effective, which is concerning given Michigan's overall low student achievement on state assessments (Chetty et al., 2014; Darling-Hammond, 2000; Michigan Department of Education, 2015). Many other U.S. states face similar problems, with overwhelming majorities of teachers receiving effective or highly effective ratings, making it difficult to identify which teachers are truly high-quality (Weisberg, Sexton, Mulhern, & Keeling, 2009). In 2009, U.S. President Barack Obama signed into the law the American Recovery and Reinvestment Act (ARRA), designed to support many U.S sectors, including education (U.S. Department of Education, 2009). Embedded within ARRA was the Race to the Top (RTTT) initiative, a competitive grant encouraging U.S. states to focus on reforming aspects of their education systems, including adopting more rigorous academic standards, using student data to inform and improve instruction, hiring and retaining highly-qualified teachers, and reducing achievement gaps within schools (U.S. Department of Education, 2009). RTTT also encouraged these states to change the process of evaluating teachers, with a focus on evaluations that can better distinguish teacher performance as well as provide better information on what makes a high-quality teacher (U.S. Department of Education, 2009). As a result, since 2009 more than two-thirds of U.S. states have made significant changes to how teachers are evaluated (The Center for Public Education, 2013). This nationwide reform effort has resulted in new, multi-tiered, and complex teacher evaluation policies.

Past research shows principals play a crucial role in how policies, including teacher evaluation policies, are implemented (Halverson, Kelley, & Kimball, 2004; Rigby, 2015; Spillane, Reiser, & Reimer, 2002). How principals interpret, communicate, and enact new

teacher evaluation policies has great implications for how these policies play out in practice. Because principals play such an influential role in the implementation of teacher evaluation policies, it is important to understand what impacts their sensemaking of these policies. While principals have always held some responsibility for evaluating their staff, the process is now higher-stakes because in most cases a teacher's evaluation score is tied to career defining decisions, such as hiring, firing, and tenure decisions (Harris, Ingle, & Rutledge, 2014).

In this paper I ask a) what factors influence principals' sensemaking of changing teacher evaluation policies and b) how these factors influence both decision-making by principals, as well as the ways the policies are implemented. To help answer these questions, I draw on data from interviews from six public school principals in the U.S. state of Michigan. Additionally, at the time these data were collected teacher evaluations were legislated at the local or school district level. The U.S., nor the state of Michigan, regulates exactly how teachers' are evaluated, only that teachers must be evaluated. The specifics of this evaluation process is determined at a local level. To help better understand these local level evaluation systems I collected school district teacher evaluation documents, including specific district teacher evaluation policies, teacher evaluation rubrics, and final teacher evaluation ratings.. Based on an analysis of the findings, I argue the high-stakes nature of teacher evaluations shaped principals' sensemaking and implementation of these policies. Specifically, because new teacher evaluation policies were tied to the employment of their teachers, principals made sense of these policies in very specific ways.

Literature Review

This study builds on the line of research that suggests principal sensemaking influences individual interpretation and implementation of school level policies. While early research based in the U.S. suggested principals lacked the power and influence to change school and teacher practices (Bidwell, 2001), more recent research suggests principals play a key role in reform and policy implementation (Coburn, 2005; Donaldson & Papay, 2014; Gawlik, 2015; Halverson et al., 2004; Rigby, 2015; Spillane, 2006; Spillane, Diamond, Burch, Hallett, Jita, & Zoltners, 2002; Spillane & Kenney, 2012). One specific policy that principals have made sense of (in some form) for the past century is how to evaluate their teaching staff. Early research suggested principals in the U.S. played a more hands-off role when evaluating their staff, rarely evaluating teacher instruction and instead evaluating teachers on things such as if they showed up to work on time (Darling-Hammond, Wise, & Pease, 1983; Hallinger, Heck, & Murphy, 2014; Wise, Darling-Hammond, McLaughlin, & Berstein, 1985).

However, as U.S. teacher evaluation policies began to transition into more high-stakes policies, the role of the principal transitioned as well. Principals were tasked with evaluating classroom instruction and providing feedback to teachers, taking on the dual role of coach and evaluator (Duke & Stiggins, 1986, 1990). More recent work has focused on the role of the principal as an "instructional leader", where the principal is charged with supporting teacher instruction and is held accountable, along with teachers, for student learning (Blasé, Blasé, & Phillips, 2010; Smylie, 2010). Currently, as most U.S. schools use rigorous teacher evaluation systems, which typically include a student achievement based component, as well as an observational component with a detailed and structured observation rubric, the role of the principal in the evaluation process is much better defined than in previous years. For example, in most situations principals are given specific directions by state and/or district level authorities of how and when to observe teachers, how to score teachers, and how to provide feedback to teachers (Goldring, Grissom, Ruben, Neumerski, Cannata, Drake, & Schuermann, 2015). State and local authorities also tell principals how to account for and use student assessment data in a teacher's final evaluation score and how they are to account for other non-

instruction teacher actions (i.e., mentoring other teachers, teacher attendance at professional developments, etc., Goldring et al., 2015).

Despite the more detailed and structured role given to current school principals, research shows even the most diligent school leader practitioners and policy implementers adjust their sensemaking based on their local context and the meaning they give to a policy (Coburn, 2005; Halverson et al., 2004; Spillane et al., 2002). Weatherley and Lipsky (1977) stressed the importance of "street-level bureaucrats" – the individuals who impact how policies are ultimately implemented. These individuals and their cognition, including their beliefs, skill, will, resources, time, context, and capacity, impact how policies look in practice (Halverson & Clifford, 2006; Kimball & Milanowski, 2009; Weatherly & Lipsky, 1977). In this paper I argue principals are a specific type of street-level bureaucrat, controlling how teacher evaluation policies play out in practice.

In sum, while principals are experiencing more clarity and structure around how they are to evaluate their teaching staff, principal cognition and sensemaking still greatly impacts how these policies look in practice. As the role of U.S. principals evaluating teachers has evolved in recent years, more research is needed to better understand how principals think about this more defined role. This line of research is particularly important as the quality of principal leadership is a strong predictor of student achievement and teacher retention and satisfaction (Boyd, Grossman, Ing, Lankford, Loeb, & Wyckoff, 2011; Clark, Martorell, & Rockoff, 2009; Grissom, 2011; Leithwood, Seashore-Louis, Anderson, & Wahlstrom, 2004; Seashore-Louis, Wahlstrom, Leithwood, & Anderson, 2010).

Theoretical Framing: A Sensemaking Perspective

Sensemaking theory is a constructivist assumption that acknowledges learning is shaped from past experiences and prior knowledge and that learning occurs through our social and situational context (Weick, 1995). A growing body of literature in education uses sensemaking theory to address questions of how people and organizations interpret and implement school policies (Coburn, 2005; Halverson et al., 2004; Koyama, 2014; Rigby, 2015; Spillane et al., 2002). Research on school principals that uses a sensemaking theory lens provides two broad strands of findings; (1) principals' prior experiences greatly influence how they understand and make sense of new policies (Harris et al., 2014; Jacob & Lefgren, 2008; Nelson, Sassi, & Grant, 2001); and (2) local context impacts how principals make sense of and implement policies (Coburn, 2005; Spillane et al., 2002). Despite these broad strands of findings research currently lacks of more nuanced understanding of how school principals implement evolving teacher evaluation policies and systems. This study aims to provide more information on what factors influence principal sensemaking of teacher evaluation policies (besides experience and local context), as well as provide information on how these factors influence how principals make human capital and evaluative decisions about their teachers.

The sensemaking frame emphasizes individuals bring in set experiences, beliefs, values, and perceptions each time they are introduced to some event and these experiences, beliefs, values, and perceptions influence how individuals interpret and implement reforms and policies (Fullan, 2001; Weick, 1995). This personal sensemaking process helps explain how and why individuals charged with a task interpret and ultimately implement this task. While earlier research suggested policy design and local resistance may contribute to unsuccessful implementation of education policies more recent work suggests the implementation process is much more complex. As Spillane et al. (2006) write, "This work (research in education policy implementation) suggests that viewing implementation failure exclusively as a result of poor clarity or deliberate attempts to ignore or sabotage policy neglects the complexity of the human sensemaking processes consequential to

implementation" (p. 47). In this study, I use a sensemaking theory lens to analyze what factors influence principals' interpretation and implementation of evolving teacher evaluation policies.

Study Context: Educator Evaluations in the U.S. State of Michigan

The changes the U.S. state of Michigan has made to its teacher evaluation policies are consistent with what many other states are doing across the U.S., which makes it a timely and interesting case to study. Michigan's effort at reforming teacher evaluation laws began in 2010 with legislative talks about reworking teacher tenure laws. This effort culminated in Public Act 101 in 2011, which increased the probationary period of beginning teachers from four years to five years and legislated that an untenured teacher, if rated effective or highly effective, could not be removed from their current teaching placement solely based on seniority (Michigan Department of Education, 2015). While these changes began Michigan's teacher evaluation reform effort, a big shift occurred with the creation of the Michigan Council on Educator Effectiveness (MCEE) in 2011. MCEE consisted of educational researchers, educational experts, school principals, and members of the Michigan Department of Education (MDE) and was charged with developing a fair, rigorous, and transparent evaluation system for evaluating teachers and administrators. Together, these educational experts spent 18 months reviewing the most recent research across the United States and globe regarding the most effective and fair way to evaluate teachers. In June of 2013, MCEE submitted their final proposal to overhaul Michigan's teacher evaluation system. At the time these data were collected, districts were encouraged to adopt one of four teacher evaluation observation rubrics (5 Dimensions of Teaching and Learning, Charlotte Danielson's Framework for Teaching, Marzano Teacher Evaluation Model, or the Thoughtful Classroom) and districts were encouraged to base 25 percent of a teacher's final evaluation score on student assessment data.

In qualitative research it is important the researcher have some familiarity with the setting and phenomena under study (Miles, Huberman, & Saldaña, 2014). I have five years' worth of experience working in U.S. public schools, first as a teacher and then on the administrative side as an instructional coach. During my four years as a teacher our school district used three different teacher evaluation systems. My time as an instructional coach was largely spent introducing teachers to another new teacher evaluation system adopted by our school district. I experienced this myriad of teacher evaluation reforms from the perspective of a teacher and from the perspective as an administrator, resulting in my interest in how administrators think about and ultimately enact teacher evaluation policies and systems. As a current Ph.D. candidate in educational policy I have focused extensively on consuming research on teacher evaluation policies and systems and how school leaders make sense of teacher evaluation policies and systems. My intentions with this work (and my scholarly agenda as a whole) is to provide nuanced information on how principals make sense of teacher evaluation policies and how this sensemaking process influences how principals make decisions. I believe this is an important topic because in the U.S. teacher evaluation policies are very high-stakes and are used to for teacher career-defining decisions, such as hiring, firing, pay, and retention decisions. In most cases, school principals are the primary decision makers with regards to a teacher's final evaluation score and my goal is to better understand what goes into these high-stakes, complex, and extremely important decisions school principals are asked to make.

Research Methods

To address this study's research questions I chose an exploratory case study approach. Exploratory case studies are particularly useful when attempting to answer "how" and "why"

questions regarding a particular phenomenon (Yin, 2014). Additionally, as Wynsberghe and Khan (2007) write:

> We propose that case study could be considered a transparadigmatic and transdisciplinary heuristic that involves the careful delineation of the phenomena for which evidence is being collected (event, concept, program, process, etc.). Our definition of case study has led us to suggest that case study is not exclusively about the case revealing itself as it is about the unit of analysis being discovered or constructed. This is an important development because it means that researchers cannot definitively state the unit of analysis at the outset of the research; it must come into focus as the research progresses. (p. 90)

Given this definition, an exploratory case study approach has the potential to shed light on the phenomenon of how school principals think about and evaluate teachers. Six public school principals in the U.S. state of Michigan participated in this study. Participants were selected using criteria-based sampling (Creswell, 2013). Participants had to meet each of four criteria for selection: (1) all participants had to be located in the same U.S. state, as teacher evaluation policies vary from state to state; (2) all participants had to work in public schools that were under the same state teacher evaluation policy requirements; (3) participants were required to have a minimum of five years of experience as a principal, in an effort to include principals who had experience implementing previous teacher evaluation systems; and (4) participants must have adopted the same teacher evaluation system, in order to eliminate some of the variation caused by including principals who used differing teacher evaluation systems. Three participants in this study worked in elementary schools, one worked in a middle school, and two worked in high schools. The participants' experience as a school principal ranged from six years to more than 20 years. All participants in this study used the Charlotte Danielson Framework for Teaching (Danielson) and all participants factored in 25 percent student assessment data into their teachers' final evaluation score.

I relied on two sources of information for this study: (1) one-on-one in-person interviews with school principals and (2) district teacher evaluation artifacts. All participants in this study were interviewed in the fall of 2014, as they were in the middle of evaluating their teaching staff. The purpose of the semi-structured interviews was to better understand how principals were making sense of evolving teacher evaluation policies. Questions focused on principals' knowledge and beliefs about their current teacher evaluation system, actions principals took throughout the process of evaluation, and the factors the principals attribute to influencing how they think about and enact teacher evaluations. The interviews, each lasting between 45 and 60 minutes, were conducted over a period of four months and were audio-recorded and transcribed. Additionally, district-specific teacher evaluation policies were collected from all participants. The author of this study collected all of these data. The purpose of collecting these documents was to better understand what principals were being asked to do by their district and to compare these documents to the interview data, focusing on the alignment and consistency of what principals' were being asked to do by state and district requirements and what they were actually doing in practice.

The author of this study input all data into Atlas.ti, a qualitative data analysis software. The author used this software to analyze and interpret patterns, trends, commonalities, and links among the participants (Miles, Huberman, & Saldaña, 2014). The author coded all of the data inductively and as themes emerged they were grouped together (Miles, Huberman, & Saldaña, 1994). The author then reviewed all code groupings looking for common excerpts that highlighted similar themes and ideas. After completing this coding, the author checked the validity of the coding process by recoding the data for a second time. The author noted any

discrepancies and these discrepancies were addressed in order to refine and justify assertions and to look for possible other alternative interpretations of the data (Miles, Huberman, & Saldaña, 2014). Once the codes were identified the author grouped together common stories, themes, and ideas generalizing within the case. After the coding process was complete, the author compared to the quotations original interview text, making sure they were taken in context. The author then took the groups of quotations and organized the quotations, as similarities emerged. To establish validity the author contacted all participants to clarify any questions that arose during the transcribing and coding of the data. Additionally, the author reached out to the participants of this study and asked them to clarify their statements before drawing any conclusions. Finally, the author received Institutional Review Board (IRB) for the Protection of Human Subjects approval to conduct this research.

Findings

The findings from this study suggest the high-stakes nature of teacher evaluations shaped principals' sensemaking and implementation of their district's teacher evaluation policies. Specifically, because new teacher evaluation policies tied teacher evaluation scores to the employment of teachers, principals made sense of and implemented these policies in very specific ways. Two prominent themes emerged after an analysis of the data. First, principals overwhelmingly said they lacked confidence in their districts' teacher evaluation policy, both in terms of the rubric that was used to observe teachers and in terms of the student growth measure that accounted for part of the teachers' final evaluation score. Principals said they did not feel their districts' teacher evaluation policy was an accurate representation of teacher effectiveness or teacher quality and believed the policy was largely subjective. As a result, principals relied on their own beliefs about what constituted effective teaching, not necessarily what the policy dictated. Second, principals said they lacked confidence their peers were evaluating teachers accurately. This belief resulted in principals scoring their own teachers more favorably than what might be expected, because principals knew these scores would be compared across districts when making employment decisions.

Lack of Confidence in the Policy

All principals in this study lacked some degree of confidence in their district's teacher evaluation observation rubric. Part of this lack of confidence was due to the principal's perception that these observation rubrics were largely subjective. While all principals in this study agreed new teacher evaluation measures were more objective than previous systems (in terms of evidence-based examples of good instruction), each principal mentioned how they, as the primary evaluators, still ultimately controlled the final outcome of the evaluation. For example, Dr. Little said:

> Oftentimes as a principal or some other evaluator, a lot of it is intuitive. It's somewhat subjective honestly. A lot of it's subjective. The more we try to make it scientific, it's still kind of an art. I mean, even my own observations, depending on what's going on in the rest of the building, or my level of energy that day, or the teacher's level of planning, I mean, I think those are all things that can vary so greatly. There's so many varying factors.

For Dr. Little, who has spent more than 20 years as a principal, while new evaluation policies and measures may try to take the subjectivity out of the evaluations, the principal is still the person who decides how a teacher will be evaluated. Ms. Georgio underscored this

sentiment and commented, "We still rely far too much just on our gut. Although I'd like to tell you that my gut's always right, that'd probably be a lie."

Additionally, principals unanimously agreed there was too much gray area in the observation rubric and, as a result, scoring teachers was difficult. Ms. Walker said, "I think that there is a gray area there with how many bullet points or how many check marks a teacher should receive to make them a highly effective, an effective or an ineffective teacher. That part is gray." Ms. Georgio added, "I think it's easy to see that these are really good people, driven people, great people, but to help quantify and help them take step A to B to C, it's (the evaluation system) not good enough, not even close." The belief that their district's evaluation rubrics had room for subjectivity afforded the principals an opportunity to rate teachers higher or lower than might be expected by the individuals who designed the observational rubric. Because of the principals' lack of confidence in the fairness and accuracy of their evaluation system the principals in this study did not want to rate teachers harshly. This resulted in principals defaulting to higher than might be expected evaluation scores. In short, because principals knew these evaluation scores would be used for future teacher employment, and because they were not completely confident in the system, they did not feel comfortable critically evaluating teachers.

Because of their lack of confidence in their district's teacher evaluation policy, instead of using these observation rubrics as an objective tool for measuring teacher effectiveness, principals drew on their own understanding of what teacher quality and effectiveness looked like. Because Ms. Georgio and other principals in the study were confident in their ability to accurately assess teacher performance and because of the principals' lack of confidence with their teacher evaluation observation rubric, the principals in this study believed their subjective evaluations were a more accurate gauge of their teachers' performance than their district's teacher evaluation observation rubric.

Principals unanimously agreed they did not need a more structured rubric to tell them who their effective and ineffective teachers were. For example, Dr. Little noted, "I think again, oftentimes any evaluation system is only as good as the person doing the evaluation, doing the evaluating. You have to understand the situation and make decisions accordingly." Dr. Little went on to add that no one specific system can understand the intricacies of individual schools and situations, so when evaluating his teachers, Dr. Little puts more stock in what he knows, not what the rubric dictates. He prefers to rely on his expertise because, as he notes, "You're affecting someone's life and their livelihood—at least potentially."

Another area in which principals lacked confidence with their district's teacher evaluation policy was with how their individual districts used student assessment data for human capital decisions. For example, Dr. Allen did not believe it was fair to use the same percentage of student growth data for all teachers. He said:

> This is the most at-risk building in the district with a very high at-risk population. We look very different than our colleagues across the district that have a very low percent of at-risk kids. I mean, there's just so many factors my kids come to school with that those kids don't, and so I feel like my teachers have a much more complex job than the teachers that serve in that school.

Dr. Allen was reluctant to put too much stock in the results of his school's student assessment data when comparing to schools within his district because of the unique challenges faced by his school. However, he was quick to note that the way the district's teacher evaluation system was currently constructed, all teachers' evaluation scores would be compared across the district when making layoff decisions. Dr. Allen expressed discontent for this approach for

using teacher evaluations because in his opinion comparing these student data between schools was unfair to this teachers given their unique demographic context.

Their principals lacked trust in the specific student assessments that were used for teacher evaluations. Specifically, principals felt these assessments were not an accurate representation of how much students learned throughout the school year and as a result they were hesitant to put too much stock in the results. Mr. Lin explained:

> Who knows what the reliability of those assessments are? Other assessments that are highly designed are assessments that we don't get results back for six, seven months, so those become ineffective in this. It's very difficult to say that it's fair because you don't know what you're getting.

Mr. Davidson agreed explaining, "There's some real challenges there. Is the assessment that we're using to measure their learning valid? Is it reliable? Do we trust it to really measure what's happening?"

Mr. Lin and Mr. Davidson provide two examples principals' lack of confidence in the assessments that were being used to a part of a teacher's evaluation score. These principals each said that using assessments to help teachers improve their instruction would be a preferred approach, but in their opinion using what are unreliable and flawed assessments for the high-stakes purposes of teacher hiring and firing was not a good approach to teacher evaluations. This belief resulted in a lack of confidence in their district's teacher evaluation policy because these policies valued the results of these assessments – in all cases these results were 25 percent of a teacher's final evaluation score.

The principals in this study believed the ways their teacher evaluation system used student assessment data and the actual assessments that were used to determine student growth did not do an adequate job gauging teacher performance or effectiveness. This resulted in principals further inflating teacher evaluation ratings on the observational part of a teacher's evaluation to combat this flawed system of student growth data.

Perhaps most pointedly, Dr. Little said:

> You're basing on one test perhaps given over one or two days. At that point in time, I hate to base everyone else's career on that. At that point in time, it could very, very well be career changing. I think there's a lot of things to be said— there's a lot of learning that happens in school besides some things that can be measured by a one-sitting test. It's a high-stakes, and I get it, but I don't agree with it. Philosophically, I don't agree with it.

Dr. Little provides a representative example of all principals in this study who lacked confidence in their district's teacher evaluation policy. Whether the principals lacked confidence in the observational rubric used to evaluate teacher instruction or in the assessments used to determine student growth, the principals believe their district had a flawed system of evaluating teacher performance, resulting in these principals typically scoring teachers higher than might be expected if they believed in the fairness and accuracy of their policy/system.

Lack of Confidence in their Peers

Another primary factor influencing principals' interpretation and implementation of teacher evaluation policies was concern surrounding the ways teacher evaluation scores would be compared across the district. For example, Dr. Little said:

> I think the fear piece is the biggest thing. Especially in a larger district where, for example, we have several elementary schools and one principal scores tougher than someone else. That is a concern because what happens if I'm in Principal A's building and that person scores pretty easily, and then I go to Principal B's, and he is tough. I could be minimally effective in one and effective or highly effective in another with a comparable skill set. Then it comes time if we come down to a layoff situation, the one with the lower scores will be laid off.

Dr. Little went on to articulate that although he had been trained to use the new evaluation system, he did not feel confident in this ability to make the fine-tuned decisions that may impact a teacher's future. Thus, when faced with a decision about rating a teacher, he defaulted to a higher score. This lack of confidence in himself led Dr. Little to believe that other principals must feel the same. This resulted in Dr. Little not trusting the accuracy of the teacher evaluation scores provided by his peers, because he know how he typically dealt with assigning evaluation scores.

Mr. Lin's experience was similar. He explained:

> I think the biggest problem with this process is that you think about how many hundreds of schools, I mean, that we have, thousands, really, in the state of Michigan, and every school is essentially implementing it differently. When you start looking at a district like ours, it becomes a big problem because we have 30,000 students and a lot of schools, over 100 administrators, and every administrator who's doing those evaluations is putting their subjective spin onto that somewhat objective tool. I think that creates potentials for very significant problems. For example, I have teachers who have no problem being rated effective from me but are concerned that their jobs may someday be judged against somebody who's being rated highly effective from an administrator that they don't believe is doing it right.

The principals quoted about illustrate a common sentiment of the broader group principals in this study. The principals in this study overwhelmingly felt their district's teacher evaluation system was not being implemented consistently throughout their school district. In general these principals believed most principals rated teachers higher than might be expected, because these principals did not want their teachers let go if their district experienced layoffs. Because of this belief, the principals in this study too rated their teachers more favorable, resulting in inflated teacher evaluation scores. The principals in this study constantly mentioned how their teachers' evaluation rating would be used in future employment decisions. Principal thinking about how teacher evaluation scores would be used for future teacher employment impacted principal decision making surrounding how to provide an evaluation score and as a result, principals typically scored teachers more favorably than principals would have if these scores were not used for employment decisions.

Discussion

The principals in this study used their own thinking and beliefs to evaluate teachers, and based their justification for this choice on the high-stakes nature of teacher evaluation policies.. They rationalized this decision by thinking they wanted to protect teachers from being dismissed by a system that in their opinion was not an accurate reflection of teacher effectiveness.. When analyzed through the lens of sensemaking theory, these findings suggest

that the ways a principal values or perceives the purposes of teacher evaluations, and the relationships he or she has with staff, shape how he or she interprets and ultimately implement teacher evaluation policies. As Kraft and Gilmore (2016) write, "In practice, districts often hope to promote teacher development while also using evaluations for high-stakes accountability" (p. 711). Because these principals' school district were using teacher evaluations for human capital decisions, principals were less likely to rate teachers critically. Specifically, the principals in this study interpreted and implemented teacher evaluations while always thinking of the future employment of their teaching staff. Recent work from Grissom and Loeb (2016) produced similar findings in which principals were more likely to rate teachers higher on high-stakes evaluations than on low-stakes assessments.

Prior research suggests that teacher evaluation policies are intended to perform two primary functions: teacher improvement and teacher accountability (Hanushek & Rivkin, 2010; Kraft & Gilmour, 2015; Steinberg & Donaldson, 2016). A logical question, then, is, is it reasonable to charge principals with achieving these goals? Since the RTTT initiative in 2009, changes to teacher evaluation systems have occurred across the U.S. Principals are the primary individuals tasked with understanding these changes and are expected to successfully implement these evolving, complex systems. In light of the heavy workload shouldered by the average U.S. principal (Horng, Klasik, & Loeb, 2010) expecting these individuals to add to this workload is problematic. New teacher evaluation systems task principals with rating teachers accurately and differentiating amongst teacher effectiveness, while also supporting teacher instructional improvement. Teacher evaluation systems are time consuming to implement and this implementation must be done delicately given the enormous stakes attached to current teacher evaluations policies.

Before discussing the implications of these findings, it is important to note the limitations of this study. First, although the aim of this paper is not to generalize its findings, one limitation is from sampling. The six principals in this study shape these findings by their experiences, values, and knowledge; the findings in this paper, then, could be different if six other principals were involved. Second, this paper did not observe principals interacting with the evaluation process in a naturalistic setting and therefore the reports from principals have been filtered through their interpretations and understandings. Finally, while district-specific teacher evaluation policies were collected, completed principal observation and evaluation data was not received for this study. In this way, the principals' recollection of the process of evaluating teachers needs to be taken at their word and cannot be compared to what they actually did in practice.

Despite the before mentioned limitations the data collected in this study provides insight into how the principals in this study navigate teacher evaluation policy implementation. While not generalizable to the entire principal community, the results and analysis of this work have the potential to hold in other contexts with other principals. Principals across the United States are continually implementing evolving, high-stakes teacher evaluation policies and therefore future research should be able to compare the results of this work to principals across the state of Michigan as well other contexts across the United States that are experiencing similar teacher evaluation reforms.

The results of the analysis have implications for both policymakers and practitioners. First, for policymakers, this analysis suggests factors outside of, but related to, teacher evaluation policies (i.e., future teacher employment) impact how principals make sense of and ultimately implement teacher evaluation policies. The principals in this study constantly referred back to the idea that a teacher evaluation was a snapshot in time and did not encapsulate all that teachers did during the school year. Additionally, principals noted that their teacher evaluation systems had room to adjust final scores if they felt such adjustments were called for (e.g., if a principal felt the outcomes of teacher observations and/or students' final

assessment data did not reflect teachers' true impact on student learning). These findings lend some credence to research that suggests using multiple observers, or observers who know little to nothing about the teachers they are evaluating, may provide more reliable assessments of teacher instructional ability (Donaldson & Papay, 2014; Kane, McCaffrey, Miller, & Staiger, 2013). The present study suggests that, because principals may feel protective of teachers' employment and future teaching career it is difficult for principals to evaluate teachers objectively. Using outside observers or multiple observers who do not know the teachers as intimately and/or who are less concerned about how the evaluations will be used has the potential to alleviate this concern. This is something policymakers and district leaders should consider when designing future teacher evaluation policies. This removes the relational aspect of performance evaluations, which remains a concern when trying to cultivate a fair and objective teacher evaluation system. Admittedly, this does not remove the ethical question surrounding the ways teacher evaluations will be used, which may be a concern for any evaluator.

Implications of these findings for practice include providing principals more structured and intensive training on how to best use new teacher evaluation systems may help principals feel more confident in the accuracy and fairness of these systems. For example, increasing evidence suggests ongoing conferences between principals and teachers are crucial to the overall teacher evaluation process (Steinberg & Donaldson, 2016; Steinberg & Sartain, 2015; Taylor & Tyler, 2012). Therefore, principals should receive constant support as to how to structure these conferences, what to include during conversations in these conferences, and how to deliver useful feedback to teachers. Principals will likely continue to play an active role in negotiating policies and initiatives (Ganon-Shilon & Schechter, 2017; Koyama, 2014), but if policymakers and district leaders could ensure that the essential parts of teacher evaluation policies were prioritized in training and ongoing support of principals, some of this lack of continuity in policy implementation discussed in this paper may be abated.

The results of this study, coupled with other emerging work (Donaldson & Papay, 2014; Goldring et al., 2015; Grissom & Loeb, 2016), may be one explanation as to why teacher evaluation policy implementation remains a challenge in the U.S. state of Michigan and beyond. Interestingly since the reform of teacher tenure laws in 2011, of the 96,000 K-12 teachers in Michigan, only 19 have been dismissed due to poor evaluation scores (Michigan Department of Education, 2015). Additionally, K-12 teachers in Michigan continue to be rated overwhelming effective or highly effective; 97% of teachers in the state met this criteria (Michigan Department of Education, 2015). According to the findings of this study, we can likely attribute these high teacher evaluation ratings and lack of dismissals due to principals scoring teachers higher than would be expected – not necessarily because all of these teachers are effective or highly effective in the classroom.

References

Bidwell, C. E. (2001). Analyzing schools as organizations: Long-term permanence and short-term change. *Sociology of Education, 74,* 100-114.

Blasé, R., Blasé, J., & Phillips, D. Y. (2010). *Handbook of school improvement: How high-performing principals create high-performing schools.* Thousand Oaks, CA: Corwin Press.

Boyd, D., Grossman, P., Ing, 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.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review,*

104(9), 2633-2679.

Clark, D., Martorell, P., & Rockoff, J. E. (2009). *School principals and school performance. CALDER: Working paper 38.* Washington, DC: The Urban Institute.

Coburn, C. E. (2005). Shaping teacher sensemaking: School leaders and the enactment of reading policy. *Education Policy, 19*(3), 476-509.

Creswell, J. (2013). *Qualitative inquiry and research design* (3rd ed.) Los Angeles, CA: Sage.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Educational Policy Analysis Archives, 8*(1), 1-44.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research, 53*(3), 285–328.

Donaldson, M. L., & Papay, J. P. (2014). Teacher evaluation reform: Policy lessons from school principals. *Principal's Research Review, 9*(5), 1-8.

Duke, D. L., & Stiggins, R. J. (1986). *Teacher evaluation: Five keys to growth*. Washington, DC: National Educational Association.

Duke, D. L., & Stiggins, R. J. (1990). Beyond minimum competence: Evaluation for professional development. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 116-132). Newbury Park, CA: Corwin Press.

Fullan, M. (2001). *The new meaning of educational change* (3rd ed.). New York, NY: Teachers College Press.

Ganon-Shilon, S., & Schechter, C. (2017). Making sense of school leaders' sense-making. *Educational Management Administration & Leadership, 45*(4), 682–698.

Gawlik, M. A. (2015). Shared sense-making: How charter school leaders ascribe meaning to accountability. *Journal of Educational Administration, 53*(3), 393-415.

Goldring, E., Grissom, J. A., Ruben, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher, 44*(2), 96-104.

Grissom, J. A. (2011). Can good principals keep teachers in disadvantaged schools? Linking principal effectiveness to teacher satisfaction and turnover in hard-to-staff environments. *Teachers College Record, 113*(11), 2552-2585.

Grissom, J. A., & Loeb, S. (2016). Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low – and high-stakes environment. *Education Finance and Policy*, 1-53.

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

Halverson, R., & Clifford, M. (2006). Evaluation in the wild: A distributed cognitive perspective teacher assessment. *Educational Administration Quarterly, 42*(4), 578-619.

Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. In C. Miskel & W. Hoy (Eds.), *Theory and research in educational administration* (pp. 66-90). Greenwich, CT: Information Age Press.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267–271.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal, 51*(1), 73-112.

Horng, E. L., Klasik, D., Loeb, S. (2010). Principal's time use and school effectiveness.

American Journal of Education, 116(4), 491-523.

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., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Seattle, WA: The Bill and Melinda Gates Foundation.

Koyama, J. (2014). Principals as bricoleurs: Making sense and making do in an era of accountability. *Educational Administration Quarterly, 50*(2), 279-304.

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

Leithwood, K., Seashore-Louis, K., Anderson, S., & Wahlstrom, K. (2004). How leadership influences student learning. New York, NY: The Wallace Foundation. Retrieved from http://www.wallacefoundation.org/knowledge-center/Pages/How-Leadership-Influences-Student-Learning.aspx

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making with a standards-based evaluation system. *Education Administration Quarterly, 45*(1), 34-70.

Nelson, B. S., Sassi, A., & Grant, C. M. (2001, April). *The role of educational leaders in instructional reform: Striking a balance between cognitive and organizational perspectives.* Paper presented at the American Educational Research Association Conference, Seattle, Washington.

Michigan Department of Education. (2015). *Educator evaluations and effectiveness in Michigan.* Retrieved from www.michigan.gov/educatorevaluations

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: An expanded sourcebook* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Rigby, J. (2015). Principals' sensemaking and enactment of teacher evaluation. *Journal of Educational Administration, 53*(3), 374-392.

Seashore-Louis, K., Wahlstrom, K. L., Leithwood, K., & Anderson, S. E. (2010). *Investigating the links to improved student learning.* Retrieved from http://www.wallacefoundation.org/knowledge-center/school-leadership/key-research/Pages/Investigating-the-Links-to-Improved-Student-Learning.aspx

Smylie, M. A. (2010). *Continuous school improvement.* Thousand Oaks, CA: Corwin Press.

Spillane, J. P., & Kenney, A. (2012). School administration in a changing education sector: The U.S. experience. *Journal of Educational Administration, 50*(5), 541-561.

Spillane, J. P., Riser, B. J., & Gomez, L. M. (2006). Policy implementation and cognition: The role of human, social, and distributed cognition in framing policy implementation. In M. I. Honig (Ed.), *New directions in educational policy implementation: Confronting complexity* (pp. 47-64). Albany, NY: SUNY Press.

Spillane, J. P., Diamond, J. B., Burch, P., Hallett, T., Jita, L., & Zoltners, J. (2002). Managing in the middle: School leaders and the enactment of accountability policy. *Educational Policy, 16*(5), 731-762.

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

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.

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

The Center for Public Education. (2013). *Trends in teacher evaluation.* Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/

U.S. Department of Education. (2009). *Race to The Top program executive summary.* Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Weatherly, R., & Lipsky, M. (1977). Street-level bureaucrats and institutional innovation: Implementing special education reform. *Harvard Educational Review*, *47*(2), 171–197.

Weick, K. E. (1995). *Sensemaking in organizations.* Thousand Oaks, CA: Sage Publications.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York, NY: The New Teacher Project.

Wise, A. E., Darling-Hammond L., McLaughlin, M. W., & Berstein, H. T. (1985). Teacher evaluation: A study of effective practices. *The Elementary School Journal 86*(1), 60-121.

Yin, R. K. (2014). *Case study research: Design and method* (5th ed.). Thousand Oaks, CA: Sage Publications.

Author Note

David is a 4th year educational policy doctoral student at Michigan State University. His current research interests include how teachers and school administrators make sense of the policies they are asked to implement. Correspondence regarding this article can be addressed directly to: reiddav1@msu.edu.

Article Citation