

SCHOOL INSPECTION IN THE UNITED STATES:
POTENTIAL FOR SCHOOL REFORM AND LASTING INSTITUTIONAL CHANGE

By

Pablo Bezem

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Education Policy–Doctor of Philosophy

2021

ABSTRACT

SCHOOL INSPECTION IN THE UNITED STATES: POTENTIAL FOR SCHOOL REFORM AND LASTING INSTITUTIONAL CHANGE

By

Pablo Bezem

In an era of test-based accountability, school inspection may offer a nuanced understanding of school performance and actionable information for improvement. Yet, little empirical evidence exists on its effectiveness in advancing performance, particularly in the United States. This dissertation examines the potential of school inspection as an alternative accountability mechanism through a series of studies based in one of the only U.S. districts to experiment with inspection. Three papers evaluate: 1) How inspectors arrive at their decisions, 2) whether inspections promote principals' attitudes that are associated with lasting institutional change, and 3) effects of inspection on school planning directions. Each of these papers offers a significant contribution to the literature on school inspection and provides evidence regarding the inspection process and its potential to enable school improvement.

The first paper inquires about how inspectors evaluate school and reach determinations. To shed light on inspectors' decision-making processes, the case of a U.S. district is contrasted with two long-established international systems. Results reveal that decisions are strongly influenced by local culture and professional traditions. Despite efforts to introduce alternative means for school assessment in the U.S., a test-based accountability mindset dominates and limits the potential of inspection.

The second paper investigates whether principals' attitudes towards inspection are those associated with lasting institutional change. This study bridges insights from the organizational change literature in the education field. Semi-structured interviews with 20 principals in the selected U.S. district inquire about the perceived effectiveness of the diagnosis, appropriateness of inspection

feedback, and readiness for implementing changes. Results show that strongly positive attitudes toward inspection associated with lasting change. A majority of principals view inspection favorably since its breadth and depth contributes to a more accurate diagnosis of key challenges. Holistic evaluation and actionable findings are not feasible through test-based accountability alone. Principals also express a strong commitment to implementing changes based on inspection feedback. These results provide the first empirical evidence of the influence of school inspection on sustained institutional change.

The third essay examines the effect of inspection on school planning. Prior to implementation of reforms, priorities are set through the school planning process. Despite the wide use of inspection globally, no previous study has tested whether a causal relationship exists between inspection and school planning. This study uses mixed methods to examine whether inspection shifts the areas of focus in school planning documents. The study sample comprises 160 public schools in the selected U.S. district. In-depth interviews with school principals reveal that inspections are perceived as useful and led to planned reforms focused in two areas: instructional practices and school climate. Results from a difference-in-differences analysis suggest that inspection shifted the focus of planning documents towards these two areas. Inspection led to nearly a doubling of keywords in improvement plans related to these two focus areas. This study provides empirical evidence regarding the potential of inspection to inform school planning.

Overall, this dissertation advances understanding of the potential of school inspection to offer insights for improved school reforms, in a high-stakes, test-based accountability system. Findings demonstrate how local and professional culture influence and condition inspection practices. In addition, I show that principals demonstrate strong positive attitudes towards inspection, which are favorable to sustainable change. Finally, I provide causal evidence that inspection shapes the focus of school improvement plans. These results have implications for U.S. accountability policies and the potential for a more comprehensive approach, beyond test-based accountability.

With all my love to Maurita—my dearest partner in this journey—Aidan—my new inspiration—
and Hernán & Alicia—the bedrock for all my endeavors—.

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	ix
Introduction	1
Paper 1: Informing New Approaches to School Accountability in the United States: Inspectors’ Decision-making from a Comparative Perspective	4
Introduction	4
Background: U.S. Education Accountability	6
Literature Review: Inspectors’ Decision Making	7
Theoretical Framework	9
Methods	10
Case Selection	11
International Comparison	11
U.S. District	11
Rio Negro, Argentina	12
The Netherlands	13
Data Collection and Analysis	14
Results	17
Indicators of School Quality	18
U.S. Case: Focusing on Rubrics and Avoiding Bias	18
Comparison: Role of Indicators of School Quality	20
Multi-Informant Approach	22
U.S. case: Finding Trends and Discarding Outliers	22
Comparison: Role of Multi-Informant Approach	23
Interactions Among Inspectors	25
U.S. Case: Consensus Building & Guided Sensemaking	25
Comparison: Role of Interactions among Inspectors	28
Local Context Information	29
U.S. Case: Minor Role in Inspectors’ Thinking	29
Comparison: Role of Local Context Information	30
Inspectors’ Perspectives	31
U.S. Case: Personal Judgement within the Scope of the Protocol	31
Comparison: Role of Inspectors’ Perspectives	34
Conclusions and Policy Implications	37
The Legacy of Test-Based Accountability in the United States	38
APPENDIX	40
REFERENCES	48
Paper 2: Principals’ Attitudes towards School Inspection in a U.S. district: Contribution to Sustained School Reform	60
Introduction	60

Background	62
Literature Review	64
Theoretical Framework	66
Case Study: Description of School Inspection System	68
Methods	69
Results	72
Perceptions about the District Diagnosis Effectiveness excluding Inspection	73
School Inspections: Positive Attitudes among Principals	75
Perceptions of Diagnosis Effectiveness	75
Sentiments of Appropriateness	78
Readiness for Change	80
Mixed Sentiments and Ambivalence	82
Concerns about Diagnosis Effectiveness	82
Concerns about Appropriateness	83
Uncertainty about Readiness for Change	84
Negative Attitudes	86
Conclusions	88
APPENDICES	92
Appendix A: Interview Protocol	93
Appendix B: Coding Scheme	97
REFERENCES	99
Paper 3: The Effect of Inspection on School Improvement Planning: Evidence from a US. District	111
Introduction	111
Literature Review	113
School Change based on Inspection Feedback	113
The Uses of School Improvement Plans	114
District Background	116
School Inspections	116
School Improvement Plans.	117
Research Design	118
Stage I. Interviews with School Principals	119
Influential areas based on inspection feedback	120
Stage II. Content Analysis	121
Word Frequencies in SIPs - Most Influential Areas	124
Stage III - Statistical Analysis	125
Results	128
Interview Analysis - Perceived Usefulness of Inspections	128
Difference-in-Differences Analysis	133
Conclusions	138
APPENDICES	142
Appendix A – Codebook for Interviews to School Principals	143
Appendix B – Content Analysis Dictionary	144
Appendix C –Most Frequent Phrases on School Improvement Plans – School Years 2016-17 and 2018-19	146

Appendix D – Test Score Results	147
Appendix E – Dictionary for Placebo Test	148
REFERENCES	149

LIST OF TABLES

Table 1. Inspectors' Background and Experience	15
Table 2. Sources Guiding Inspectors' Thinking	16
Table 3. Principals' Experience and Education	69
Table 4. Principals' Attitudes towards School Inspection	72
Table 5. Summary of principals' views	76
Table 6. Influential Areas - Changes Implemented in Schools based on Inspection Feedback	121
Table 7. Content Analysis Coverage and Term Frequency	125
Table 8. Summary Statistics	127
Table 9. Non-parametric Difference-in-Difference	133
Table 10. Content Analysis Coverage for Panel	133
Table 11. DD Regression Results	135
Table 12. Balance Tests	137
Table 13. Placebo Tests.	138

LIST OF FIGURES

Figure 1. Comparison: Influence of Information Sources on Inspectors' Thinking	18
Figure 2. Research Design	119
Figure 3. Parallel trends: Word Count for Inspected vs. Not Inspected Schools	136

Introduction

Test-based accountability (TBA) prevails as the central paradigm for school improvement efforts across the United States (e.g. Figlio & Loeb, 2011; Hanushek & Raymond, 2005). Schools are incentivized to raise student achievement on standardized tests (e.g. Ladd & Figlio, 2008). Yet, test scores alone offer limited insight into specific reforms that might benefit a given school (e.g. Gagnon & Schneider, 2019). An alternative approach to accountability is school inspection, which is widely used outside of the United States. Instead of primarily relying on test scores, inspection consists of holistic evaluations in schools and are conducted by expert educators. These evaluations include classroom observations, school document reviews, and interviews with school staff, students, and families. By closely observing school operations, inspectors might gain better insight into factors that might help or hinder improvement (e.g. Barber, 2005).

Despite the promise of inspection to enable reforms, there is little empirical evidence on its effectiveness (de Wolf & Janssens, 2007; Ehren, 2016b; Klerks, 2012). No prior empirical studies have assessed inspection in the United States. This dissertation examines the potential of school inspection as an alternative accountability mechanism by developing a case study of one of the only U.S. districts to experiment with inspection. Through three papers, different aspects of inspections are evaluated: 1) Inspectors' decision making from an international comparative perspective, 2) Principal attitudes towards inspection associated with lasting institutional change, and 3) Effects of inspection on school planning. Each of these papers offers a significant contribution to the literature on school inspection and provides evidence regarding the potential of inspection to enable school improvement.

The first paper focuses on school inspectors' decision-making and the role of sensemaking in their evaluations. Using a comparative case study of a U.S. district and two long-established international systems, this study examines the information sources that guide inspectors' thinking. It

assesses their use of professional judgement and the implications of inspectors' sensemaking for school improvement. Results show that decisions are strongly influenced by local and professional culture. In the U.S. case, a "test-based accountability mindset" dominates the inspection process. This leads to inspections strictly adhering to protocols and reduces inspectors' professional insights in an effort to avoid bias. In contrast, in the two international cases, inspectors rely more on their professional judgment and delve into complex issues beyond the limited procedures outlined in protocols. Despite efforts to introduce alternative means for school assessment in the United States, the prevalent mindset might limit the potential of inspection to gain insights for school improvement.

The second paper investigates whether principals' attitudes towards inspection are those associated with lasting institutional change. This study is grounded in organizational change theory and uses semi-structured interviews with principals. Responses are compared between inspected and not inspected schools in the selected U.S. district. Interviews inquire about perceived effectiveness of the diagnosis, appropriateness of inspection feedback, and readiness for making changes. Results show that strongly positive attitudes toward inspection lead to dispositions that are associated with lasting change. A majority of principals highlight that the breadth and depth of inspection contribute to a more accurate diagnosis of school' problems. Such a holistic evaluation with actionable findings is not available through test-based accountability alone. Principals also express a strong commitment to making changes based on the inspection feedback. These results provide the first empirical evidence of the effects of school inspection in enabling sustained institutional change.

The third paper uses mixed methods to examine the causal effect of inspection on the focus areas in school planning. No prior study provides evidence of the causal effect of inspection on school planning. A step prior to implementing school reforms is typically setting priorities through the school planning process (e.g. Matthews & Sammons, 2004). In-depth interviews with school principal reveal that inspections were generally perceived as useful for planning purposes and led to anticipated or actual reforms. Inspection particularly influenced two areas of reform: instructional

practices and school climate conducive to learning. Next, the study evaluates the influence of inform on these two areas of reform, using content analysis and difference-in-differences analysis. Content analysis evaluated the presence of terms related to these areas in school improvement plans. A difference-in-differences analysis finds that inspection shifted the focus of planning documents. Inspection led to nearly a doubling of key words in school improvement plans in related to these focus areas.

This dissertation reveals several aspects of inspection relevant to inform policymaking: contrasting global practices, implementing sustained reforms, and the effect of inspection in shaping school planning. The first paper offers a contrast between the U.S. case and long-established inspection systems. It shows that even when there are apparent similarities between formal inspection mechanisms, culture influences practice and decision-making. It reveals a trade-off between strict adherence to the protocols to avoid bias versus greater flexibility and reliance on professional judgement. It shows that the potential to gain insights from inspections in the U.S. system may be constrained by the professional culture. Despite this evidence, the second paper shows that the principals in the U.S. case perceive that the larger scope of inspection offers a more accurate and actionable diagnosis than the TBA system. Further, most principals demonstrate strong positive attitudes that are associated with lasting institutional change. The third paper builds causal evidence that inspections led to actual changes in the school improvement plans.

Overall, this dissertation sheds light on the potential of school inspection to offer insights for improved school reforms. As the implementation of Every Student Success Act sparked debate over the design of more comprehensive accountability designs, inspection could be viewed as an alternative or complement to test-based accountability. Given the lack of a consistent body of literature on holistic, on-site evaluations of schools in the United States, this dissertation contributes to this debate by shedding light on the potential and limitations of school inspection.

Paper 1: Informing New Approaches to School Accountability in the United States: Inspectors' Decision-making from a Comparative Perspective ¹

Introduction

High stakes testing prevails as the central paradigm for school improvement efforts across the United States. This is driven, in part, by an audit culture that emphasizes performance measurement as a primary policy focus (Apple, 2005; Clarke & Ozga, 2011). An appealing feature of test-based accountability (TBA) is the perception that it objectively measures educational performance and thus allows comparisons across districts and years (cf. Bloem, 2015). Yet, TBA typically does not provide nuanced information to identify *why* certain schools fall behind; it does not capture the myriad factors that influence school quality (e.g. Darling-Hammond et al., 2016; Gagnon & Schneider, 2019). As a stand-alone policy, TBA often incentivizes schools to narrowly focus on tested subject areas (e.g. Fitchett & Heafner, 2010; Jacob, 2005) and strategies to boost scores, which might not promote substantive learning (Rothstein, Jacobsen, & Wilder, 2008).

School inspection (SI) is an alternative approach for monitoring and improving school quality, that emerged in national policy discussions on how to redesign accountability systems after the 2015 enactment of the Every Student Success Act (Darling-Hammond et al., 2016; Klein, 2016; Ladd, 2016; K. Ryan et al., 2013). Unlike TBA, school inspection is not limited to standardized tests to evaluate student performance. Inspection evaluations usually assess a variety of school processes through observation and direct contact with school stakeholders, such as teachers, students, and parents (van Bruggen, 2010). In this way, inspections can provide a summative assessment of overall

¹ This paper was led by Pablo Bezem and co-authored with Dr. Anne Piezunka and Dr. Rebecca Jacobsen.

school quality while also uncovering factors that help or hinder school improvement. Several U.S. districts have experimented with inspection in some form including New York City, Los Angeles, Oakland, and Cleveland. In the U.S., inspections are often referred to as Quality Reviews. Lessons from previous experience with inspection can inform the debate about redesigning the U.S. accountability system.

School inspections might serve as an alternative mechanism to achieve more nuanced accountability and insights for improvements. Yet, a major concern is its perceived subjective nature (Glazerman, 2016), which can raise questions about its reliability. A tradeoff exists between obtaining insights from inspectors with subjective perspectives versus achieving greater reliability with standardized metrics. To understand how inspection works, it is critical to shed light on inspectors' decision process and how school quality is judged. No previous study has evaluated what drives inspectors' decision-making in the U.S.; internationally, scarce empirical literature exists regarding this aspect of SI.

While TBA relies on evaluation of academic content knowledge to determine school quality, SI depends on the complex decision-making process of individual inspectors. In this way, inspectors are the linchpin for SI reliability and effectiveness. Understanding inspectors' thinking and how personal perspective are utilized can shed light on this process. This can then inform policy discussions regarding the advantages and limitations of SI, compared to TBA. This study poses three research questions: (1) What are the sources guiding inspectors' thinking during inspections?, (2) How do personal perspectives of inspectors influence school evaluations?, (3) What implications does inspectors' sensemaking have for school improvement and school quality?

To address these questions, we develop a comparative case study of inspection in a U.S. school district and two long-established international SI systems. Our case in the U.S. is one of the few districts in the country to consistently use inspection for over a decade. The two international cases include a province in Argentina (Rio Negro) and the Netherlands. The Argentinean case has a

system of continuous support, flexible procedures, and low stakes; the Dutch case has a high-stakes system with a formal protocol and expert-based flexibility for inspectors. These three systems represent distinct contexts and can illuminate evaluation processes and inspectors' decision making.

Background: U.S. Education Accountability

The United States has assembled one of the most developed education accountability systems in the world (Figlio & Loeb, 2011). This system has been shaped by New Public Management principles in the 1980s that promoted greater rationalization, evidence-based change, output orientation, and rigorous accountability. The emphasis on accountability in policy reform was evident in the 1990s, with some states implementing high-stakes accountability systems. These became widespread with the passage of the No Child Left Behind (NCLB) Act of 2001, which mandated the nationwide use of test scores to measure of school quality. This legislation led to a new definition of school reform that was broadly supported by elective representative across the political spectrum (Figlio & Loeb, 2011; Ravitch, 2016) and became common sense among school reformers. This situation has been described as the *audit culture* (Apple, 2005).

Subsequently, calls for a broader set of quality indicators and state-level flexibility led to the passage of the Every Student Succeeds Act in 2015. This sparked a nationwide discussion regarding how to redesign U.S. accountability systems. Already, most states and districts use a greater variety of school quality metrics (Edgerton, 2019; Portz & Beauchamp, 2020). Although the U.S. continues to emphasize TBA (Mathis & Trujillo, 2016), greater flexibility for states and districts creates new possibilities for more holistic approaches for accountability and school improvement (Darling-Hammond et al., 2016). Inspection mechanisms emerged as an alternative approach to evaluate school quality (Darling-Hammond et al., 2016; Klein, 2016; Ladd, 2016, 2017; K. Ryan et al., 2013). SI considers a greater variety of factors that influence education quality, such as inputs, expert observation of school process, and interaction with school stakeholders.

Literature Review: Inspectors' Decision Making

Policy efforts to incorporate inspection mechanisms into U.S. accountability systems are limited by scarce empirical evidence (de Wolf & Janssens, 2007; Ehren, 2016b; Klerks, 2012). Despite the fact that inspection systems have long existed around the world, most previous research focuses narrowly on European systems. Furthermore, the wide variety of SI arrangements that exist makes it challenging to build a coherent body of literature that converges on key findings. The local nature of inspection has reinforced a tendency of focusing inspection research on local systems, which is often then published in country-specific journals in the native language. SI research published in more widely read journals, in the English language, has expanded during the last five years. Yet, the literature remains limited and fragmented. The empirical literature that does exist has tended to focus on the effects and side effects of SI (e.g. Altrichter & Kemethofer, 2015; Klerks, 2012).

Despite this growing body of research, limited empirical research has centered on school inspectors themselves and their influence on the evaluation process. Most early studies published in English were conducted primarily in the UK, where the inspectorate, the Office for Standards in Education, Children's Services and Skills (OFSTED), is a longstanding institution. Despite using a highly standardized inspection procedure and a reliable system of classroom observation (Peter Matthews et al., 1998), various studies conclude that the professional judgement of OFSTED inspectors played a key role in their evaluations (Gilroy & Wilcox, 1997; Lee & Fitz, 1997; Woods & Jeffrey, 1998). It was found that inspectors' feedback to schools is influenced by perceived constraints that the local context imposes on teachers (Woods & Jeffrey, 1998). In addition, inspectors' professional background impacts their judgement. For example, prior experience serving as a classroom teacher can increase empathy and a sense of collegiality with teachers (Baxter, 2013; Millett & Johnson, 1998).

Since professional judgement plays a role in inspectors' decision making, it is relevant to explore how individual judgement varies. Although this question has not been directly studied, Silcock and Wyness (1998) shed some light on the issue, finding a wide diversity in inspector beliefs about education and current reforms as well as their empathy with challenges faced by teachers. These differences were apparent despite standardized training and evaluation tools. This early research demonstrated that profound differences in core beliefs regarding education can persist in a highly standardized system. However, it does not attempt to link how these beliefs influence evaluations and school feedback.

Recent studies have focused on the process of judgement formation in SI systems, where feedback is decided through consensus among a group of inspectors (Dedering & Sowada, 2017; Lindgren, 2015; Rutz et al., 2017). Despite the use of protocols and standards, inspectors have some discretion, both as individuals and the overall group, when working towards a consensus and making decisions (Dedering & Sowada, 2017; Rutz et al., 2017). Lindgren (2015) demonstrates that in the highly standardized Swedish system, there is a stark contrast between how decisions are formed during the inspection process (the "backstage" of inspection) versus how final feedback is presented to the school and community (the "front stage"). Even when inspectors present hard evidence to justify decisions in the "front stage," there is negotiation among inspectors in the "backstage," where their judgments encompass a mix of uncertainty, adaptation, and creativity.

These findings show that the human element and professional judgement remain central in the inspection process, regardless of efforts to standardize processes and procedures. Despite knowing that variation does occur, it is not yet understood *how* specific personal aspects of the inspectors and institutional features of the school system influence the inspection process. This study aims to provide initial insights into this critical aspect of inspection systems.

Theoretical Framework

This study draws on sensemaking theory as a conceptual framework (Weick, 1995). A growing body of literature in education draws on this theory to understand teachers' and administrators' interpretative frameworks when enacting educational policies (Coburn, 2005; Halverson et al., 2004; Rigby, 2015; Spillane et al., 2002). Sensemaking theory is particularly useful for understanding how individual actors comprehend a situation, make meaning of it, and then act based on this interpretation (Weick, 1995; Weick et al., 2005). Educational studies that draw on this approach have addressed how this process is influenced by preexisting worldviews, prior knowledge, experience, formal and informal networks, and the organizational and social context within which sense-makers work (Ball & Bowe, 1992; Coburn, 2001; Hill, 2001; Porac et al., 1989; Spillane et al., 2002).

Sensemaking literature related to policy implementation has focused on how knowledge structures are accessed and applied in practical situations. One finding is that observations made by individuals who implement policy can often focus on the superficial aspects of a situation that then trigger a memory of another situation. This jeopardizes the ability to dive into the deeper significance of what is observed (Spillane et al., 2006). This literature has also found that individual reasoning about a complex judgement tends to be biased toward interpretations that are consistent with their beliefs and values (Spillane et al., 2002). Research has also found that sensemaking processes are mediated by considerations about organizational structures (e.g. work environment, norms, and rules), professional affiliation and networks, and traditions (e.g. Coburn, 2001; Spillane, 1999; Spillane et al., 2006). Policy implementation studies have shown the relevance of socially mediated sensemaking. For example, when teachers implement instructional policies, sensemaking is mediated by school leaders' participation in the interpretation of the policies (Coburn, 2005) as well as by interactions with other teachers (Coburn, 2001; Hill, 2001).

A separate body of literature in organizational studies focuses on sensemaking within organizations. Sensemaking theory has been used in this field to understand confusing or ambiguous events within organizations (Maitlis & Christianson, 2014; Sandberg & Tsoukas, 2015; Weick, 1995). Similar to sensemaking research on education policy implementation, organizational studies highlight the importance of constructing intersubjective meaning, which occurs when various actors within an organization, such as managers and peers, shape each other's understanding (Gioia et al., 1994).

Using these perspectives and building upon past research, our study draws on sensemaking theory to understand how inspectors interpret situations that they observe in schools and arrive at judgements regarding school quality. School inspectors must reconcile government guidelines, best practices, and inspection protocols with the situations they find in the schools. Therefore, sensemaking theory provides a useful lens to understand this process. Sensemaking is likely mediated by inspectors' own experience and beliefs about education, the interaction with other inspectors, and organizational culture. The sensemaking literature provides useful constructs to capture the variety of factors that influence how inspectors reconcile policies with practice (e.g. Coburn, 2005; Hill, 2001). While protocols do exist, there is flexibility for inspectors to use professional judgement (Dedering & Sowada, 2017; Gilroy & Wilcox, 1997; Lindgren, 2015). Therefore, we rely on the sensemaking literature to analyze how inspectors interpret complex situations observed at schools and how they arrive at decisions.

Methods

To investigate school inspectors' decision-making process, this study uses a comparative, multi-site case study approach. A district in the U.S. serves as the main focus. We then conduct a horizontal examination of the decision process of inspectors across sites (Phillips & Schweisfurth, 2014; Vavrus & Bartlett, 2016). This comparison highlights contrasts and similarities to the U.S. case, where SI experience is limited. Through these cross-site comparisons, we characterize a

diversity of SI arrangements and practices, which can advance understanding of the broad spectrum of inspection thinking processes (Chabbott & Elliott, 2003). The other two cases provide a broader view of various aspects of the inspectors' decision-making process, showing similarities and differences with the U.S. and identifying aspects of inspection not captured by the U.S. case. The analysis focuses mostly on inspectors' thinking processes. It takes into consideration less formal aspects of these processes, including inspectors' personal perspectives, such as preferences, beliefs, and professional judgement.

Case Selection

We selected one district in the U.S. since it is one of the few cases in the country to consistently use inspecting mechanisms for over a decade. We then contrast this case with two long-established international SI systems, Argentina and the Netherlands. Within all of three SI systems, a group of experts conducts in-school evaluations and use several modes of data collection – classroom observation, school stakeholder interviews, and document analysis. Yet, these cases also differ from one another in key aspects relevant to our study objectives. Notable differences include the purpose that inspection serves (accountability vs. support) and severity of consequences due to the inspection results. Differences in protocol of conducting inspections are also present – the frequency and length of inspection visits, number of inspectors, and public availability of inspection reports. These differences may influence the information sources that inspectors consider when evaluating schools.

International Comparison

U.S. District. The U.S. case is a large, urban school district that relies heavily on an high-stakes framework, based on high-stakes test scores, for accountability purposes. The district began experimenting with inspection processes more than a decade ago as part of school reforms. The inspection program, referred to as *Quality Reviews*, primarily targets low-performing schools. Unlike the other cases included in this study, inspection is outsourced to private consulting firms and is not

directly managed by a governmental office. Since 2012, the process has been led by a company we will refer to as *QualiEv*. This inspection program gathers qualitative evidence about school programs for accountability and formative reviews. School visits are conducted by groups of three to four inspectors. At least one is a representative from *QualiEv*, a full-time inspector who leads the process, and the others are certified reviewers from the District Department of Education. The team is guided by a detailed protocol which outlines the evaluation process and includes research-based standards regarding effective school practices. Inspection activities consist of school document reviews, classroom observations, as well as interviews and focus groups with teachers and administrators. Immediately after inspection visits, inspectors share main findings in an oral report to school administrators. Then, inspectors and school administrators work jointly in a planning process, discussing school strengths and areas of growth, establishing the next steps, defining strategies, setting measures to establish success, and a timeline to achieve these goals. A written report summarizing conclusions is provided to schools, which includes suggestions for priority areas, but not specific recommendations for improvement.

Rio Negro, Argentina. Each province in Argentina manages its educational system. Inspection is the main mechanism for school accountability. Standardized tests are low-stakes and only used for diagnostic purposes. Inspectors are in the highest position of the teaching ladder. They are full-time public officials who report directly to the provincial Ministry of Education and must have considerable experience: 12 years in teaching, 2 years in leadership, and inspectorate training (Concurso de Supervisores Rio Negro, 2013).

The main purpose of SI in Rio Negro Province is to provide support to all schools. In the process, school administrators are held accountable. Inspectors develop inspection projects, and while they must follow broad guidelines, there are no specific protocols for school visits or inspection activities. Inspectors are assigned to a group of schools to conduct administrative controls and provide continuous support. School visits occur at least three times a year, and can be more

frequent if a school requires more support (Resolución Del Consejo Provincial de Educación de Río Negro N 1053, 1994). Inspectors consult with their technical team of professionals in education to inform their work. All inspectors go through basic training and are accountable for following legal standards. Inspectors prepare reports for the schools, which are not publicly available. No sanctions are imposed for poor academic performance. Furthermore, inspection does not track standardized educational outcomes, such as test scores, nor must follow specific standards regarding education processes.

The Netherlands. The Dutch Ministry of Education coordinates educational policy with municipalities. Accountability relies on both outcome- and school-based components (Nusche et al., 2014). High-stakes testing has an important role (Scheerens et al., 2012; van der Sluis et al., 2017), while at the same time, inspection is a central instrument for monitoring standards. Inspectors are full-time public officials and receive specific training.

The Netherlands emphasizes SI of low-performing schools.² While all schools are inspected at least once every four years, the lowest performing receive more frequent and rigorous visits. To determine the frequency and type of inspections, inspectors use a risk-based model. This model assesses school risk based on administrative information, including standardized test scores, accountability documents, and failure signals, such parents' complaints or negative media reports (Education Inspectorate - Ministry of Education, 2010). Inspectors follow an assessment framework for legal aspects, process quality, and outcomes. The framework pays particular attention to learning outcomes, educational process, school environment, quality assurance and ambition, and financial management (Education Inspectorate - Ministry of Education, 2017a, 2017b). Each of these areas include a set of standards that is operationalized based on statutory requirements. Results of

² At the time of the interviews, the Dutch system was transitioning to School Board inspections in addition to continuing with the risk-oriented school inspections. In our interviews, we focused on the on-site school inspections as implemented until the academic year 2016-17.

inspection are shared with the school and public through a summary report. If schools do not demonstrate improvement for two years, inspectors can recommend to the Ministry of Education administrative and/or funding sanctions. In the most extreme cases, this can lead to school closure (Ehren,Altrichter,McNamara, & O'Hara, 2013; OECD,2015).

Data Collection and Analysis

We conducted semi-structured interviews with inspectors of K-12 schools in the three study locations. We inquired about the inspector's background, activities performed during the inspection process, and outcomes of inspection. In addition, we asked about how they make decisions regarding school quality and what aspects of quality they value the most. Emphasis was placed on capturing the inspectors' thought process through the use of probes that asked for concrete examples to illustrate their thinking and we provided inspectors with scenarios to gauge how they would response to a given situation.

In the United State, we interviewed inspectors from the district Department of Education who were certified by *QualiEv*. We invited all 29 certified reviewers who had previously conducted inspections. For the other sites, we selected a purposive sample of inspectors (Teddlie & Yu, 2007). The objective was to conduct 6 to 10 interviews at each site. In total, we completed 23 interviews: 8 in the United States, 9 in Argentina, and 6 in the Netherlands. Interviews lasted an average of 72 minutes. In the United States and Argentina, interviews were conducted in-person in the local language, English and Spanish, respectively. In the Netherlands, interviews were conducted via videoconference in English. Interviews in Argentina were transcribed in Spanish and then translated into English. Interviews from the Netherlands and the U.S. sites were transcribed and checked for accuracy.

Participants were informed that interview responses were anonymous, transcripts would not be shared, and a pseudonym would be used to cite them. In the United States and Argentina, inspectors were given a US\$ 25 gift card after participation. In the Netherlands, we were advised by

SI researchers not to offer incentives. Participants in each study site are in-line with the characteristics of inspectors in their location with respect to years of experience and demographics. Descriptive information about interviewees are presented in Table 1.

Table 1. Inspectors' Background and Experience

Variable	U.S. Case n=8	Netherlands Case n=6	Argentina Case n=9
<u>Individual inspectors</u>			
Inspector experience, in years	2.3	9.8	6.2
Education experience, in years	14.3	22.8	31.0
Classroom teaching experience, in years ¹	8.9	10.7	13.1
Administrative experience, in years ¹	2.7	7.0	15.2
<u>% of inspectors, within case</u>			
% former classroom teachers	100%	50%	100%
% former administrators	38%	20%	100%
% with graduate degree	100%	75%	56%

¹ Only those inspectors with experience as teachers/administrators were included in these indicators.

The interview transcripts were coded using deductive and inductive codes. Deductive codes were formulated based on our theoretical framework, mainly from concepts related to sensemaking theory (Coburn, 2005; Maitlis & Christianson, 2014; Spillane et al., 2002, 2006; Weick, 1995). Inductive codes stemmed from interviews in the three sites. Responses were coded according to Miles, Huberman, and Saldaña's (2014) approach to qualitative analysis by observing patterns and themes within and across case studies. We used DEDOOSE qualitative data analysis software for coding and analysis. To ensure the reliability of codes, we used an independent-coder method. First,

interview transcripts were independently coded by two researchers, before then comparing the coding for agreement. We followed an iterative process until at least 75% agreement was achieved.

We conducted two rounds of coding. The first round focused on: i) sources of information used during inspection, ii) use of local context information, and iii) inspectors’ definition of good quality education. These codes were defined inductively. The second round relied more heavily on deductive codes that were more abstract and required more interpretation. The main codes include iv) inspectors’ perception of school administrators, v) type of recommendations, and vi) sources guiding thinking. This latter code is emphasized in our analysis (Table 2). The sub-codes are based on the inspection procedures and sensemaking theory. The theory was used to define the foci on the knowledge structures accessed by inspectors when facing practical situations, especially when they have to make sense of complex situations.

Table 2. Sources Guiding Inspectors’ Thinking

	Source	Description
A	Indicators of school quality	Standardized rubrics, indicators, or metrics used to evaluate school quality during inspection
B	Multi-informant approach	Simultaneous use of multiple sources of information - of the same or different kind - to validate evidence
C	Interactions among inspectors	Interaction among inspectors or between inspectors and technical personal to discuss about findings
D	Local context information	References to local context information, including students' demographics, characteristics of the neighborhood, school history, and change of staff
E	Inspectors’ perspectives	References personal experience, beliefs, or professional judgment

After the coding procedure was complete, we identified general and country-specific patterns in the data using descriptors in DEDOOSE. The data were examined visually in the form of code

clouds, cross-tabulations, and charts showing the frequency with which codes occurred as well as the presence or absence of codes within and across interviews, for each of the case studies. To ensure that reported patterns are an accurate representation of each study site, we shared initial findings with interviewees and incorporated their feedback (Miles et al., 2014).

Results

Overall, our comparative analysis indicates that personal perspectives influence inspectors' evaluations. The extent of influence is affected by local culture, professional traditions, and values. In particular, our U.S. case relies on high-stakes testing for accountability and its audit culture emphasizes performance measurement and data-driven decision making. U.S. inspectors' thinking is infused with this culture and their decision making is highly influenced by a high-stakes accountability and a standardization mindset that leads to strictly adhering to protocols and reduces inspectors' professional insights in an effort to avoid bias. Despite opportunities presented by SI to dig deeper and identify unique strengths and weaknesses at schools, U.S. inspectors actively disregarded insights that do not fit within the confines of the protocol. In contrast, the Dutch and the Argentinean cases illustrate approaches in which inspectors have more flexibility and rely on their professional background and judgement. Inspectors routinely pursued "surprising" observations even when these do not fit neatly into a protocol or the anticipated focus areas. Instead, these inspectors adopt a more holistic understanding of school quality and avenues for improvement.

In this section we present findings for the categories of sources that guide inspectors' thinking: Indicators of school quality, Multi-informant approach, Interaction among inspectors, Local context information, and Inspectors' perspectives (see definitions in Table 2). Key findings regarding the influence of these categories on inspectors' thinking across our cases are summarized in Figure 1.

Figure 1. Comparison: Influence of Information Sources on Inspectors' Thinking

	U.S. Case	Netherlands Case	Argentina Case
Indicators of school quality	High	Low	Low
Multi-informant approach	Low	High	Low
Interactions inspectors	Low	Low	Low
Local context information	Low	Low	Low
Inspectors' perspectives	Low	Low	High

Legend: Degree of Influence on Inspectors' Thinking

High	Low
------	-----

Note: The relative influence of each source was determined by number of mentions in interviews.

Indicators of School Quality

U.S. Case: Focusing on Rubrics and Avoiding Bias

In the U.S. case, indicators of school quality in the form of a standardized rubric are the cornerstone of the evaluation process and all of the inspectors mentioned them repeatedly when explaining their thought process. Their data collection is structured around various aspects of school quality, a series of metrics to measure them, guidelines for making observations, and questions for schools in order to evaluate each indicator. Unlike the two other study sites, we found that U.S. inspectors restrict data collection almost exclusively to sources specified in the protocol: classroom observations, interviews and focus groups, and school documents. Only two inspectors mentioned that they seek publicly available school information before their visit (e.g. test results and school

website) or during the visit (e.g. teachers' planning documents, students' files). This stands in stark contrast with our other two sites where developing a deep knowledge of a school before the actual visit is a vital part of the inspection process.

Moreover, in the U.S. case, the protocol explicitly determines the structure and nature of classroom observations and the questions to be asked in interviews and focus groups. For example, when inspector Amy-US was asked about what information she personally looks for or asks to see apart from the required data, she emphasized the importance of adhering to the protocol:

I just followed the ... protocol. I ask only the questions that are outlined for teachers and only the questions that are outlined for students. If a student [response] needs to be elaborated, I would say "could you tell me more ...," but I don't bring my own questions to the process or anything like that. I just tried to follow what is asked of me.

We asked inspectors about the value they placed on different information sources to evaluate school quality. We found that observing classroom instruction is consistently the most valued source of information used by inspectors to evaluate school quality. In contrast to the other study sites, most U.S. inspectors considered other sources of information to play a secondary role, if considered at all. School planning documents were deemed the least valuable source of information to evaluate school quality and were used mostly for triangulation purposes. Similarly, the use of school climate observations—such as culture, interactions among students and teachers, and facility condition—was also secondary.

More than in the other study sites, when the U.S. inspectors explained their thought process, they repeatedly made direct remarks about how they try to avoid personal bias when completing the rubric and following how the protocol defines good instruction. This was highlighted by Sarah-US and Donna-US:

Sarah-US: We all come to the table with our own expertise, with our own beliefs and values, with our own biases or preconceived notions about what school should look like ... the rubric

then helps people put those things aside, understand their influence, and then really ground themselves in both evidence and the rubric to get to a shared understanding.

Donna-US: It's really important ... as a reviewer, to not have a bias ... I might have a bias towards what good instruction looks like. So instead of using the rubric in front of me, I'm going towards what I think is good. Or I might have a bias towards what a functioning school environment looks like and sounds like. So instead of using the evidence in front of me, I'm just going towards what I think ... I think that can be both positive and negative.

Across all U.S. interviews, we found a rigid emphasis on observing the protocol, sometimes in ways that appear to impede important insights. Most inspectors recognized the relevance of their own expertise and experience; they acknowledged which indicators best capture school quality and their preferred information sources. Yet, as we show in sub-section “E) Inspectors’ perspectives,” U.S. inspectors dismissed this wealth of knowledge and equated personal views with bias. Incorporating such information into evaluations was viewed as a validity threat to the inspection process. U.S. inspectors claim to actively suppress the influence of their education knowledge and experience during inspection by stating that they strictly adhere to protocol. Furthermore, unlike the other research sites, when explaining their thinking in specific situations, they were rarely able to share concrete examples based on their own experience. Instead, they repeatedly referred to “good practices” outlined in the rubric.

Comparison: Role of Indicators of School Quality

Similar to the U.S. case, Dutch inspectors utilize indicators of school quality in all their school visits. However, these indicators were not a central aspect in inspectors’ narratives about their thinking process. While the policies and procedures used by the Dutch shares similarities with the U.S. case, and its *QualiEv* protocol, Dutch inspectors expressed greater flexibility in the inspection process, including stages of data collection, choosing which criteria to focus on and emphasize, and how to interpret indicators of school quality. When collecting information, Dutch inspectors use

standardized data only as a starting point. Our interviews revealed that data collection and usage are guided by inspectors' choices to probe more deeply into key areas, as their insights and understanding of a given school evolve over the course of the visit (We explore this further in the sub-section below, which addresses the multi-informant approach). No consensus exists among the Dutch inspectors regarding the most valuable information sources. Several inspectors found this question difficult to answer, in contrast to the U.S. inspectors who promptly referenced protocols. Some Dutch inspectors especially valued interviews with teachers and administrators as well as classroom observations. Only a few mentioned school climate or student interviews. Dutch inspectors were more likely to view the usefulness of information sources as highly contextual, based on their experience, school context, and specific issues that emerged during a school visit.

In contrast to these two cases, in Argentina, a standardized set of indicators of school quality have almost no role in the inspection process. When indicators are collected, they are mainly used for administrative purposes. Inspectors do not use standardized metrics to evaluate quality systematically. This is not to say that the inspectors lack standards for evaluating school quality. Rather, they have considerable freedom to decide what information to collect and how to evaluate schools. Consistent with this freedom, we observed immense heterogeneity in terms of information sources used and which sources are most valued. Inspectors emphasized the importance of gaining an understanding of how the school functions, emphasizing the importance of "being present" in the school, "walking around," and "living in the moments of the institutional life." Being present allows inspectors to be critical and provide support to the school. Without insights from school visits, inspectors do not feel they could truly understand a school. Therefore, they would be unable to provide necessary assistance to help schools improve. Planning documents are considered useful for evaluation and collaborative work between inspectors and schools.

Bias was rarely a concern among inspectors from Argentina and the Netherlands. In these two countries, inspectors did not hesitate to make use of their professional judgement, experience as

instructional experts, and familiarity with schools gained from multiple visits. Rather than be a cause for concern, this was viewed as being exactly what enables them to be effective. Furthermore, in the Netherlands, utilizing this type of knowledge is viewed as necessary to be a good inspector. Part of the inspector training process includes an extensive shadowing of experience inspectors, where those who are new have an opportunity to further develop and to learn to use their expertise. This is not to say inspectors in Argentina and the Netherlands do not reflect critically on their own practices and maintain a concern for integrity in the process. In both countries, nearly all inspectors expressed spontaneously in the interview a concern about being thorough in their analysis and the importance of justifying their conclusions in their feedback to schools. But attempts to standardize the process and a concern for validity and reliability were less prominent in our conversations.

Multi-Informant Approach

U.S. case: Finding Trends and Discarding Outliers

Once U.S. inspectors collect data, they are turned over to *QualiEv* where staff conduct a standardized rating process. The processed data are then used to assess trends and patterns in the school. This *multi-informant approach* seeks evidence that is confirmed repeatedly using the same type of information and then triangulated with other sources. For example, evidence from only one or two classrooms is insufficient to make a claim. Affirmation must be found in multiple classes and then triangulated with evidence from additional sources, such as interviews and school documents.

Many inspectors highlighted that this approach provides a holistic view, as Michelle-US explains:

[The multi-informant approach] really gives you that holistic view of, “Okay, this is what we saw in the classroom. This is what the teachers and students are doing.” But then, what are people actually saying about it? What are the parents saying, the students, and the staff? And how did those stories support one another, or how are they different?

Importantly, most inspectors expressed confidence in the focus on the trends, as opposed to outliers, as a reliable approach to evaluate the overall quality of the schools. Overall, most inspectors seem to

embrace this approach as an effective way to evaluate the quality of the school. As Laura-US explained:

Both times that I've done it [(the inspection)], it's been really clear, even after the first half of one day, what the trends are. It's been kind of shocking, because you think, "Oh, obviously, these classrooms [are] different than that." No, they're never different. It's all the same. It's always been really shocking how quickly you can come to what the big problem is. Usually it's actually been pretty easy to pick the top two or three, and because the schools that they pick to do these things are ... literally on fire, so it's pretty clear.

Interviews demonstrated how systematically the multi-informant approach is applied: Evidence is gathered, inspectors focus on major trends, and discard information that does not fit within these broad trends. Donna-US emphasize that they “are looking for trends and consistencies, versus an outlier of something that might strike you as wrong.” Contradictory information observed while data is being collected might rise a red flag and can help narrowing the focus. When this happens, there is some leeway for further inquiries as long as the search and sources of information to be used are part of the protocol. Yet, in contrast to the other research sites, we did not find examples of inspectors pursuing professional hunches that lead to additional question being asked, nor focusing on exceptional observations, nor conducting additional interviews that might lead to new discoveries.

Comparison: Role of Multi-Informant Approach

Inspectors in the Netherlands act as investigators. Each has considerable freedom in deciding which focus areas to emphasize during school visits. Several inspectors explained that after reviewing school information, prior to the visit, they try to anticipate the main difficulties at the school. Hypotheses are developed that they then seek to verify or disprove during the visitation day. Several inspectors noted how their expertise can assist in developing these initial assessments. They actively draw upon their prior knowledge and vast experiences with a wide array of schools to help them anticipate and hone in on issues that the school is facing. Rather than a standardized approach

that eliminates variation both between and within a school, the Dutch approach results in a dynamic process only guided by their protocols, not dictated by them. Lotte_{NETH} illustrates the questions that inspectors ask themselves during inspections:

I try to see what the most important papers are, and I read them. I try to think about what I might see in the school. I have some hypotheses in my head and I also see what I can make of the context of the school. For example... In what kind of neighborhood is it? What can I expect of the school? What's the difficult thing over there? Then I go to the school and be as open minded as possible because sometimes, when you already think you know what it will be, you will be very much surprised by what happens in the school ... I just have that in mind, somehow, but not have that on the front of my head. I just be open and see what happens during the day, but... I've got a [starting] schema in my head.

Dutch inspectors' interviews revealed that they too look for patterns and trends through a multi-informant approach. However, they also strive to identify conflicting evidence so that deeper inquiry can be made and observations and impressions during interviews can be confirmed. Rather than dismissing discrepancies or outliers, the Dutch inspectors view such findings as critical points for further investigation. Thus, inspectors use these insights to identify problems that often lurk beneath the surface. For example, Lars_{NETH} explained how he actively looks for points of disagreement:

[An important source of information is] talking to the teachers, like how they tell the [way] the school really works, how they perceive how the managements makes them work and doesn't really work, ... and do the teachers understand that vision and do they really use that vision inside their classrooms? And good thing is, we always visit the classes first, before we talk to all the people. So we can give back to the teachers and to the team leaders and to the director, what we saw in the classes. And so they can immediately give back how they perceive it.

A multi-informant approach guides thinking in both the Dutch and the U.S. cases (see Figure 1). However, there is a great difference in how information is corroborated. In the Netherlands, there is less emphasis on accumulating evidence through a rigid prescribed process, and more on finding hidden problems and testing if evidence can confirm nascent hypotheses. Thus, the process is dynamic and evolves during the visit. Inspectors determine in real time which additional documents to request, questions to ask, and aspects of classroom observations to emphasize.

The Argentinean inspectors also seek corroborating evidence, often comparing formal planning documents to actual practices. Similar to the Dutch case, and unlike the U.S. case, inconsistent findings are viewed as a critical window into key issues faced by the school. In our interviews, inspectors provided several examples of how unexpected cues during a visit can lead to additional sources. This was illustrated by Monica_{AR}:

I value walking in the schools. The fact of being present. Because face-to-face you can get to ask a new question about something specific, and you can be surprised. You can find something that you hadn't thought. ... [Sometimes you find that] the pedagogic proposals don't correspond with what you see in the visits, when you see they are not [using] the methodology they say they are applying ... If you take a child's workbook and you see mistakes in the corrections made by the teachers, or there are no corrections made by the teachers, you say: what is going on here?

Unlike the Dutch and the U.S. cases, the multi-informant approach was not stressed as a central aspect of the inspection process for most Argentinean inspectors.

Interactions Among Inspectors

U.S. Case: Consensus Building & Guided Sensemaking

In our U.S. case, data collection and synthesis are followed by a consensus building process, led by *QualiEv*. During the group discussion, *QualiEv* reinforces the previously mentioned factors in guiding thinking in order to avoid bias: focusing on the rubric and the trends, while discarding

outliers. In this phase, U.S. inspectors have an opportunity to explain their observations. *QualiEv* guides the discussion and consensus building process. Most inspectors rely on and trust in the contracted organization for facilitating the discussion and “*pushing their thinking*” (Amy-US). In this process, *QualiEv* ensures all claims are aligned with the rubric. This process was explained by Sarah-US:

So the [QualiEv] team leads a collaborative consensus building process, but they lead that in alignment with their practice and process. So it's a collaborative effort that is heavily guided by the contracting organization... So they sit as experts on how the rubric should be utilized and how things then should be scored, but they go through a process of team consensus building. Everyone presents their evidence; they do that in a group setting, and then everyone talks through it and then determines where the preponderance of evidence fits on the rubric, which then leads to the scoring process.

Inspectors must discuss evidence until arriving at a consensus regarding the evaluation. The procedure for discussion starts from the quality claims and evaluation criteria based on best practices. Then, they discuss if there is enough evidence and how to weight this evidence in order to support each claim. In this stage, they compile the collected data and the data synthesized and trends that are identified by *QualiEv*. The inspectors emphasize that any claim must be supported by evidence. This dynamic was explained by Aidan-US:

So I think the factors that usually go into play would be, “Which ones do we have the most evidence from our observations about? How strong is that evidence?” If we didn't see checks for understanding in one classroom that's not enough that we can make a priority claim around checks for understanding whatever the case might be. And so it's usually about what is the weight of the evidence that we have.

When asked whether inspectors gather further information if they have not yet found evidence to support or disprove a quality claim, Heather-US responded:

No, there's no return observation it just, the claim is tweaked based upon what you did see.

So one of the norms they [*QualiEv*] often use is see the donut, not the hole. So it's not about what you didn't see, it's about what you did see. So if you didn't see any evidence towards that claim, you go with the evidence you did see.

Avoiding personal bias is a focus during group discussions. During the formal evaluation, inspectors strive to not introduce their views regarding alternative criteria that might be informative when assessing quality. This happens during the during discussion process, as illustrated by Michelle_{US}:

When you're collaborating with a team... you share something that you saw or heard ... you're looking for another example of it. And if you don't, then you let it go. ... You don't want to be biased or making comments based on personal opinion. So, you do very much keep it factual, and you make it collaborative so it's not just one person saying one thing.

Most inspectors perceive the process of reaching a consensus to be straightforward. We did not find evidence of heated group discussions or inspectors challenging the evaluation results. Furthermore, several inspectors indicated that for many quality claims that require consensus, *QualiEv* develops preliminary statements before convening inspectors for a discussion. This appears to be a feature of the way that the process is structured. Multiple U.S. inspectors visit a given school and attend selected classes. Thus, inspectors only observe a portion of instruction; this might not encompass all domains and factors to be evaluated. For example, when we asked Aidan_{US} whether he maintains his positions when observing something that differs from the bulk of the data, he stated that he looks at "the overall picture": "I could've gone to three classrooms in the morning and in those I didn't see a particular aspect, but other people did. What I saw is just one part of all the data that's collected." This partial observation may limit the ability to develop a full view of the school and might cause inspectors to adopt the narrative of the contracting organization. This situation fits the concept of guided sensemaking in which leaders actively build a narrative that promote understandings and explanations of events (Maitlis, 2005). On the other hand, this configuration restricts individual

sensemaking and the scope of the intersubjective construction of meaning during group discussions (Gioia et al., 1994).

Comparison: Role of Interactions among Inspectors

In Argentina and the Netherlands, inspectors are employed as dedicated government staff and are placed in regional offices. In the Argentinean case, each inspector leads a team comprised of various educational expertise: Pedagogy, psychology, school administration, and social work. The group dynamic is established by each inspector and varies substantially. In some cases, the technical team actively participates in decision making and discusses which strategies are generally effective. Most inspectors use the team to make school visits in conflictive or complex situations and make specific interventions in schools. Consultations with other inspectors can be initiated; these mostly occur for complex situations or when inspectors have doubts about regulations. Some consult regularly with their colleagues in the office, others through text messages, and other in occasional provincial meetings.

In the Netherlands, inspectors receive ongoing group training and meet weekly to discuss current education issues, research and the inspection process. Some of these meetings focus on specific practices and feature invited speakers, while others offer training videos on classroom observation. When asked about how inspectors learn how to conduct inspections, Lotte^{NETH} commented on interactions among colleagues:

I'll read articles, and we've got information sessions at the office, where someone tells you something he's been working on or some interesting... with some colleagues of mine, we organize lunch sessions in which we've invited someone from outside the inspectorate to tell us and inform us about certain subjects... [for specific subjects] we try to invite someone from outside our office. We get new input and we also use our team to discuss about the standards ... "How do I interpret what I see? How do you interpret? Okay. Do we come to

the same judgments or do we judge something differently? What is the difference between us?”

Most Dutch inspectors mentioned that they generally consult with colleagues, but not for the purpose of achieving consensus. Dutch inspectors relied the least on interactions with other inspectors to inform their thinking process during an inspection. In secondary school, where inspections are conducted in groups, inspectors naturally interact with each other. However, none of the respondents emphasized a role for consensus building during the school visit. In primary school, where usually only one inspector visits the school, the process has been described as solitary, such as Emma_{NETH}: “*You go to a school alone. You arrive alone... You think alone.*” Nonetheless, some inspectors noted that they consult with colleagues in the inspection office when they face complex or challenging situations.

Interaction among inspectors plays a distinct role in the sensemaking process within each of the three cases. In the Netherlands and Argentina, inspectors work in a more stable group that share a long-term conversation about their inspection practices more generally rather than focused on a specific school example. Thus, the constant conversations shape the construction of meaning within the inspectorates (Rouleau, 2005). In the Dutch case, these conversations also have a more formal component in their ongoing training. In both cases interactions among inspectors regarding specific schools occur when they face controversial or complex situations. In this regard, inspector teams act as a sounding board to provide feedback and suggestions, but no formal consensus is required. This is in contrast to the U.S. case, where interactions are used to systematically integrate information and identify trends. Controversy and complexity tend not to be addressed.

Local Context Information

U.S. Case: Minor Role in Inspectors’ Thinking

Consideration of the “local context” plays a minor role in U.S. inspectors’ narratives about their thinking. Local context includes student demographics, neighborhood characteristics, or a

school's history. Several inspectors explained that after an extended time working in the district, they had interacted with most schools at some point and became familiar with the local context of the schools. Some of them mentioned that they have this knowledge, but they view it only as background information, as illustrated by Aidan-US who said that this contextual information is "in the back of their head." However, inspectors never mentioned this type of information as factoring into their understanding of a school's functioning. The interviews did not reveal many specific examples of how this knowledge influences thinking during inspection. Considerations of school context are not explicitly described in the formal protocol. Therefore, most inspectors actively strove to exclude this information, stating that "it *should* not matter" in their evaluations. When asked for further explanation, some inspectors highlighted that they must closely follow the rubric and avoid bias, thus assuming that objectivity of the process might be compromised by considering context. Several inspectors go further by arguing that inspected schools are low-performing and thus major differences in school context are not present.

Comparison: Role of Local Context Information

Unlike U.S. inspectors, the Argentinean and Dutch inspectors consider contextual information a critical piece of information for understanding a school's functioning. Argentina is the only research site in which each inspector is permanently assigned to a group of schools. Therefore, inspectors become deeply acquainted with the local context, student demographics, school history, and school staff. In most interviews, Argentinean inspectors highlighted that the inspections are "situated," locally oriented based on the school reality. As pointed out by Alejandra-AR: "*Everything you do in the school is based on the context, in the situational aspect, that is what I ask for, that the pedagogic project depart from there.*" To a greater extent than the other case studies, inspectors interviewed in Argentina continuously reference their knowledge of school context when interpreting problems, prioritizing information sources, interpreting student performance indicators, and determining recommendations.

In the Netherlands, inspectors also work with a fixed group of schools. Yet, after several years, inspectors switch groups as a way to ensure objectivity. Dutch inspectors exhibited more knowledge of student demographics students compared to U.S inspectors. They also offered detailed descriptions of challenges faced by schools, with specific contextual circumstances such a large immigrant population where learning Dutch was critical or parents having low educational capital to support learning at home. This local knowledge was used by inspectors to interpret various sources of information collected during inspection. For example, inspectors consider whether a school with a high proportion of immigrant students should develop provisions for language education in their planning documents. As Sven_{NETH} explained,

If you have a school with parents who speak at home another language, the schools have to invest more in curriculum in vocabulary of Dutch for those children. Then, the expectation about the quality of curriculum are different... You cannot put that into strict criteria. ... [Another example,] if you are in a small school that has to put children of several [grades together] in one group, you know it's a very hard job for the teacher to organize the lessons in a way that he challenges all the children ... so, this kind of situation plays into the way you judge the quality of instruction.

In the U.S. case, the protocols do not include context as part of their evaluation. Explicit consideration about how local context influences evaluation of school quality played a very minor role in inspectors' narratives. This configuration downplays the role of situational factors that might spark sensemaking processes in inspectors (Sandberg & Tsoukas, 2015).

Inspectors' Perspectives

U.S. Case: Personal Judgement within the Scope of the Protocol

We find that U.S. inspectors prioritize objectivity and reliability; there is an emphasis on standardized rubrics, which constrains the use of personal and professional knowledge. Yet, nearly all U.S. inspectors believe their background in education provides a necessary qualification for their

role. All interviewed inspectors had experience as classroom teachers and their work at the district Department of Education involves evaluation of classroom instruction (see Table 1). Fewer inspectors (less than half) have experience as a school administrator; among those who do, the average experience is less than three years. Some inspectors noted that an instructional background was necessary because they knew what to look for during school visits. This was illustrated by Sarah_{US}:

I think that having [an] instructional background is critically important... being an educator, someone who is highly familiar with the instructional aspect of education ... folks who have that instructional-specific lens, who carry with them the lens of what high-quality teaching and learning looks like. If you know what it means to stand up in front of students and deliver instructional content and assess students. I mean there's a lot of insider language in the rubrics. ... You have to know what you are looking for, so you have to know what teaching looks like.

In addition, three of the interviewees had experience as administrators and they believed this was important preparation for their role as an inspector. Lisa_{US} explained how her judgment is informed by administrative experience, which can provide a more systemic view:

I really think that my years as assistant principal helped me because I'm able to see the school as an entire system and not just as one specific part, and so I think that's a great qualification [for] understanding ... So if some classrooms are having anger management issues, if it's not at trend across the entire school, if there's a bigger trend arising that the instruction isn't rigorous, that that's a bigger focus for the school in trying to work with some individual classrooms.

Experience working in the district office is additional factor that inspectors feel prepares them.

Several inspectors mentioned that this experience allows them to “have a sense of what the schools look like.”

We found that in most cases, inspectors rely on their professional judgement in ways that are within the scope of the protocol, yet the approaches might not be explicitly stated in the protocol. Furthermore, in some cases, interviews revealed a tension between using professional judgement to complete the rubric and maintaining an unbiased and uniform process. This tension was illustrated by Michelle_{US} who noted how she must reconcile the evaluation rubric with the wide variety of elements she personally considers during classroom observation:

Michelle_{US}: When I'm in the classroom, I look for student engagement, and comfort, and listening, and learning. ... I think, because too often we can focus just on the teachers or the adults. ... we really have to look at the kids. When you're in a school environment, it's holistic... you're using your senses, right? You're looking, and you're feeling, and you're hearing, and all of these different things that you get when you're in a place that is not necessarily on any rubric, but you get the vibe, and the feeling of it. And then, you kind of couple that with what people are saying in the interviews, and what their body language is, and their emotional level, and how they respond to things.

Interviewer: ... how do you put all of this together in the rubric and the feedback to the schools?

Michelle_{US}: Well, those things I was just sharing, I do personally. So, those aren't necessarily on the rubric. But I think that's what comes out when you're collaborating with a team. You don't want to be biased or make comments based on personal opinion.

Michelle's comments illustrate the tension many of the U.S. inspectors expressed between valuing their professional background in education and trying to adhere to the rubrics so as not to appear biased.

Finally, we found that when crafting feedback for inspected schools, U.S. inspectors use their judgement mainly for diagnostic purposes. Unlike the other sites, we did not find many explicit considerations regarding how feedback and outcomes from the inspection affect the schools. This

shows that U.S. inspectors' sensemaking process is delimited by the scope prescribed by the protocols and is essentially retrospective more than prospective (Gioia et al., 1994; Sandberg & Tsoukas, 2015).

Comparison: Role of Inspectors' Perspectives

In direct contrast to the U.S. case, inspectors in Argentina and the Netherlands rely more heavily on their personal perspectives. In Argentina, inspectors shared openly the ways that their personal experience, beliefs and professional judgement influence multiple aspects of inspection. The value placed on this wealth of knowledge might be due to inspector's positions being the highest step in the teaching professional ladder in Argentina. More than in the other research sites, inspectors frequently made explicit remarks about how they rely on their experience to inform decisions. This was illustrated by Alejandra-AR explaining that her recommendations to schools are not only based on government norms:

Alejandra-AR: Based on what the educational policy is posing, but mixed with my perspective and stance, what I've learned all these years. Obviously, the educational policy gives you a framework in many regards. You can't stray from what is stipulated. But within these limits, my experience and knowledge are also important when the time comes to make suggestions.

Inspectors in Argentina did not shy away from frequently explaining how their personal perspectives influence their thought process. Several inspectors explained that the process is informed by their views on what they consider critical issues in education. Inspectors see their role as more political, as several explained how they act as a bridge between the macro and the micro policies of schools.

Since inspectors from Argentina determine their procedures, in contrast to the other cases, they have considerably more leeway to use their personal judgement. As in the case of the U.S. district, inspectors rely on their experience as teachers to judge teaching quality in the classroom. This was expressed, by Marcelo-AR who said that "classroom presence, [allows you] to verify the processes.... After so many years, you trust your intuitive knowledge. And you can realize very

quickly whether the kid learned or not what he should have.” But unlike the U.S. case, Argentinean inspectors’ use of professional judgement goes beyond classroom instruction and includes a wide range of aspects of institutional life, including observations of the climate, interactions among teachers, and the relationships of the schools with the families and the community.

Also, inspectors in Argentina tend to use their judgement for a wide range of issues. Since inspectors make recommendations for interventions in schools, they are obligated to go a step further and make recommendations regarding how to correct the problems identified. They must judge which practices are likely to be effective at a given school. We found that interview excerpts in Argentina coded as “personal experience and beliefs” show high co-occurrence with the parent codes “recommendations to the schools” and “responses to struggling schools.” This differs from the U.S. case, where inspectors only evaluate aspects included in the protocol and restrict their personal judgement, with the exception of classroom instruction evaluation.

In the Netherlands, inspectors’ leeway to manage and direct the inspection process within their framework, offers opportunities for relying on their personal preferences and using their professional judgement. In this process, several inspectors explained how they determine what the problems are, relying on their expertise and “gut feeling”. Some inspectors make a distinction between “hard data” found in school statistics and documents and “soft data” that is more reliant on their judgement. This was illustrated by Lars-NETH:

Some [documents are] just results, like how much of the children are on the right level when they’ve left primary school and are in secondary school now. So you can’t argue with that... you can argue about ... “how did you come to these results?” That’s the hard part, but the other parts, the soft parts, like giving chances to children... those are not always in the papers, so you can only see that in [person], when you’re at the school, and well sometimes you can get a feeling of how it should be at the school... it’s a bit of a gut feeling ...

All the Dutch inspectors provided examples of how they rely on their professional judgement to inform the process and decide on final feedback to schools. However, Dutch inspectors' narratives about their thinking process did not explicitly refer to their prior professional experience as heavily as the Argentinean inspectors.

We found that inspectors in Argentina and the Netherlands tend to provide holistic judgements of school quality and are more vocal about their personal views regarding higher-level goals of education, what is a good quality education, and how schools should function. In their narratives, their thinking is mediated by holistic judgements focused on what they believe is important for a school. In the Netherlands, for example, to evaluate the quality of the school, Lars_{NETH} asks "what is important for the kids?" and "what is the school administration doing to give the best education they can?" And Lotte_{NETH} asks how she "would feel if she had kids in the school." In Argentina, when there is a specific conflict situation in a school, Carlos_{AR} listens to students and tries to view the situation from their perspective. In the U.S. case, inspectors avoid mentioning this type of thinking, which they fear poses a risk of introducing bias.

Unlike the U.S. case, Argentinean and Dutch inspectors' narratives about their thinking not only focus on diagnosing the current situation, but also how the school has progressed and how the feedback might affect the school. Therefore, inspectors' sensemaking process is both prospective and retrospective. Familiarity with schools from previous inspections facilitates the prospective emphasis (Kaplan & Orlikowski, 2013). For example, we found that inspectors in both countries use their personal knowledge of stakeholders as an indicator of school quality. In Argentina and the Netherlands, nearly half of inspectors believe that a key indicator of school quality is their "confidence that the school administrators understand and address the main problems faced by the school," or more generally, "trust in the administrators." This was illustrated by Sven_{NETH} when he was asked how he responds when a school faces weakness, but is not failing:

I think that has to do with trust. Then you try to predict the future. You look at the quality of the staff and the quality of the management and you ask yourself the question ‘if they are not at the level they have to be at the moment, but do I trust improvement process, do I think the improvement process will go on and they will really improve, the quality education will improve in one or two years.’

The fact that inspectors from Argentina and the Netherlands highlight trust in school administrators as a key indicator of school quality might be attributable to sustained relationships between inspectors and school stakeholders. In this way, these countries differ from the U.S. case, where the inspection process is designed to avoid repeat interactions of inspectors with the same schools, as a way of enabling an objective process.

Conclusions and Policy Implications

This study focuses on school inspectors’ decision-making and the role of sensemaking in their evaluations. Not all inspection systems operate in the same way. Even when processes appear quite similar, the actual work of inspectors can be starkly different. We found that sensemaking mechanisms shape inspectors’ evaluations in different ways at each of the three study sites. Opportunities for sensemaking in the U.S. case are limited by strict adherence to the protocol and avoidance of personal bias, disregard of local context and outliers, and avoidance of complexity. The strong guidance of the contracting organization also limits the scope of socially mediated sensemaking. Inspectors do rely on experience as specialist in instruction to make sense of what they observe during evaluations of classrooms. However, this activity is limited by the scope of the rubric and do not seem to further inform the focus of the inspection. In contrast, individual and socially mediated sensemaking play a key role in the Dutch and the Argentinean cases. The evaluation process relies heavily on inspector perspectives, experience, and intuition, as well as local context information. Complexities tend to be addressed by corroborating information sources. Corroboration influences the evaluation focus. In addition, inspectors consider progress already made within the

school and the potential impact of their feedback. Therefore, the sensemaking process is not only retrospective, but also prospective.

Consistent with previous research, our study found that inspectors' personal perspectives influence evaluations in all three case studies (Dedering & Sowada, 2017; Gilroy & Wilcox, 1997; Lindgren, 2015), albeit to varying degrees. This influence depends not only on the structure of the inspection process, but also on local culture and professional traditions. In the U.S. case the audit culture aims to reduce this influence to a minimum. Yet, in the two international cases, inspectors regard their professional experience as a means to strength their evaluations.

The Legacy of Test-Based Accountability in the United States

This study informs the debate regarding how to hold schools accountable and foster improvement. School inspection is widely used around the world, yet only recently has it been adopted in the U.S. education systems. During the past forty years, the shifted towards a New Public Management approach and the paradigm of test-based accountability has forged a path dependence in educational institutions that is difficult to abandon, both within administrative structures and street-level practices (McDonnell, 2008, 2013; see also Spillane et al., 2011). Despite the potential that inspection holds for introducing a more robust way of examining school quality and fostering school improvement, the legacy of TBA continues to prevail in U.S. inspectors' thinking. Their thinking is dominated by efforts to preserve objectivity of the inspection process through strict adherence to the protocols. If something does not fit, it is discarded. The high-stakes accountability and standardization mindset influences the way inspectors think at all stages of the inspection process.

Our contrasting international cases have relevant policy implications as more U.S. districts experiment with inspection systems. The more flexible inspection models we explore capitalize on the professional expertise of the inspectors as well as the rich and detailed portrait of each school that they are able to paint using the inspection process. In the Dutch and Argentinean cases, inspectors rely on their personal perspectives and judgement in a more open way to investigate further when

they find possible cues, seek more comprehensive sources of information, and delve into the complexities of the local context. Evaluations in these cases lead to more holistic judgements of the key challenges faced by schools and therefore offer additional directions for improvement strategies. On the other hand, the heavy reliance on inspectors' professional judgement risks introducing conscious and unconscious bias. This might lead school administrators to question the fairness of inspections.

Future shifts towards alternatives such as inspection systems likely need to make explicit and ongoing efforts to account for past educational paradigms that might influence how these systems operate and perform. Overall, our results suggest that a more flexible approach can allow inspection to reach its full potential. Greater latitude for inspectors can allow this approach to accountability to better uncover underlying factors that hinder advancement in under-performing schools and to offer new insights to improve.

APPENDIX

Interview Protocol

IRB application ID#: STUDY00001267

SECTION I – INSPECTOR TRAINING AND PROCESS

Inspector Preparation, Training and Educational Experience

1. First, tell me a bit about your background in education?
 - a. How did you come to be a school inspector?
2. In general, what qualifications must someone possess (knowledge and/or skills) in order to be an inspector in [SITE]?
3. Did you receive specialized training to become an inspector?
 - i. What experiences or information did it include?
4. Given your expertise, what qualities or qualifications do you think are critical for an outstanding inspector?
5. What sort of formal interaction is there between inspectors in [SITE]?

IF YES:

- a.
 - i. How often do you interact with other inspectors?
 - ii. Around what topics or purposes do you interact?
 - iii. Do you ever seek the advice of another inspector when you are making a decision about a school?

IF NO FORMAL INTERACTION:

- b. Are there other inspectors you work with informally? How did those relationships develop?

The SI Process

6. While I've read about the SI process in [SITE]'s official documents, I know that how things work in reality can be different. So, can you describe for me the school SI process as you experience it?

FOLLOW UP AS NEEDED

- a. How often are schools inspected?
- b. How are schools assigned to inspectors?
 - i. Do you stay with schools you've inspected previously?
- c. What do you know about a school before a SI?
 - i. In your experience, are reports and information from schools at risk as trustworthy as other schools? If not, in what ways do they differ?
 - ii. How much do you know about the local context of a school – SES, race/ethnicity demographics, ELL population, other contextual challenges, etc.?
- d. How much time do you spend at a school during an inspection?
- e. Do you visit a school alone or as a team?
- f. Are there other things you look for or people you talk to beyond what's officially required?
 - i. *IF YES:* How did you select those things?
- g. What do you spend most of your time doing during a SI? (interviewing, observing classes, reviewing documents, reviewing test scores, reviewing other school performance indicators, checking safety related items, etc.)

Recommending Reform based on SI Observations

7. Are specific reform suggestions included in your final report?

IF NO – Skip to 8

IF YES

- a. How detailed are the suggestions?
 - i. What are some typical recommendations you've made?
 - ii. Do you recommend both *what* to reform and *how* to reform it?
 - iii. What are some of the most extreme recommendations you've made?
- b. How do you know what to recommend? What knowledge or experience do you draw upon?
 - i. Are there established reform strategies that have been approved by the government?
 - ii. Do you read any research on school reform?

IF NO – Skip to C

IF YES:

1. What do you read? What are your sources of research?
2. How often do you read research?
3. How do you find research on school reform?
 - a. Does anyone send it to you?
 - b. Does DPS send you research?
4. How do you then use that research in your SI process?
- iii. Do inspectors ever discuss what reforms work and don't work? In what context?

- c. To what degree are your suggested reforms co-developed with the school (leaders and/or teachers)?
 - i. IF YES CO-DEVELOP, Is there a difference in this process for a low performing schools?
 - d. When you consider sanctions or a bad report for a school, do you prefer to be more conservative with your recommendations or do you tend to propose major changes? Why do you prefer that?
8. **IF NO:** Does someone else make these recommendations? Or is this left to the school leadership to decide?

SECTION II - SI sources of information

Next I'd like to talk more about the information or data you use when making your decision about school quality.

- 9. First, can you tell me the required sources of information or data you must collect, observe or review during a SI?
- 10. What else do you personally look for or ask to see?
 - a. Why do you ask for/look at these additional things?
 - b. What do they tell you that the officially required sources don't?

IF TEST SCORES NOT MENTIONED IN THE ABOVE, ASK: What role do student test scores play in your SI process? Are these standardized tests that all schools/students take?
- 11. Thinking about all of the different sources of information about school quality that you gather during a SI, which three sources do you think are **most** informative to evaluate school quality?
 - a. Why?
 - b. If you had to rank these three, which is most important, second and third?

12. Of all of the information sources you gather during a SI, which three sources are *least* important?
- Why?
13. Are there *official guidelines* for which sources of information are supposed to be most highly valued?
- IF YES:** What?
 - Do the official guidelines match what you think as an expert reviewer?
 - Why or why not?
 - IF NO:** Are there unwritten rules amongst inspector regarding which sources of information ought to be more highly valued?
14. Are there sources of information you'd like to include in your assessment of school quality but cannot?
- What are the barriers to their inclusion? (official regulation, lack of time, lack of human resources, lack of cooperation by school staff, other)
15. Is there any type of finding that automatically leads to a poor rating *or formal report (if they have no ratings)*? For example, if a school has poor performance on their test scores, must they automatically receive a poor rating?
16. What role does the **specific context** of the school play in shaping how you interpret or value different sources of information?
- For example, if a school is located in a very poor neighborhood, does that shape how you view school test scores?
 - Are there times when some information is more important because of a school's context?

SECTION III – ACCOUNTABILITY PRESSURE

17. How is your final report circulated?
 - a. Is it publicly posted or shared with the school community?
 - b. Is there a rating that is posted somewhere?
18. Are there any benefits or rewards for a positive SI report?
19. What are the consequences if your inspection results in a poor performance/rating?
 - a. Are there any secondary consequences such as better teaching transferring to a new school or more able parents transferring their children to a new school?
 - b. How much time does the school have to address the issues raised in your report?
 - c. Do you or other inspectors return to assess progress?
20. Do you personally receive any rewards or sanctions for school performance or improvement?
 - a. *IF YES*: Can you describe what you receive and how it's determined?
 - b. Are you ever reluctant to give a negative report/rating to a school?
 - i. If yes, when?
21. I'm sure that there are clear cases of both outstanding and failing schools that are easy to spot, but I'm wondering about those schools that are on the margin. What might persuade you to give a school on the margin of failure a more positive report/rating?
 - i. How much professional judgment do you feel you can use in these cases?
 - ii. Can you give me an example of when you've been faced with a school that is right on the boarder of failure?
 1. How did you decide what to write in your report?
 2. What factors did you consider?
 3. Did you feel like you needed to leave anything out or stress anything in particular?

- iii. For those schools on the margin, is there any particular information that you find tips your decision in one direction or the other?
22. Do you rely more heavily on the government standards to make recommendations when a school is of poor quality?
- a. *If there are sanctions*: What role do potential government sanctions play in shaping your reliance on government standards?

SECTION IV - Inspector Background

23. What is your official title?
24. What educational level do you inspect?
25. How many years have you worked as a school inspector?
26. How many years total do you have working in the field of education?
27. What is your level of education? (specify if is post-secondary or university level and years of education)
28. What specific education degrees do you hold?
29. Do you have experience as a classroom teacher? How many years? What level/subject?
30. Do you have experience as a school administrator? What position/s? How many years?
31. Do you have any other education related experience?

REFERENCES

REFERENCES

- Ahuvia, A. (2001). Traditional, Interpretive, and Reception Based Content Analyses: Improving the Ability of Content Analysis to Address Issues of Pragmatic and Theoretical Concern. *Social Indicators Research*, 54, 139–172. <https://doi.org/10.1108/781350505>
- Allen, R., & Burgess, S. (2012). *How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England* (No. 12; 87).
- Altrichter, H., & Kemethofer, D. (2015). Does accountability pressure through school inspections promote school improvement? *School Effectiveness and School Improvement*, 26(1), 32–56. <https://doi.org/10.1080/09243453.2014.927369>
- Apple, M. (2005). Education, markets, and an audit culture. *Critical Quarterly*, 47(1–2), 11–29. <https://doi.org/doi:10.1111/j.0011-1562.2005.00611>
- Armenakis, A., Bernerth, J., Pitts, J., & Walker, H. (2007). Organizational Change Recipients' Beliefs Scale. *The Journal of Applied Behavioral Science*, 43(4), 481–505. <https://doi.org/10.1177/0021886307303654>
- Armenakis, A., & Harris, S. (2009). Reflections: our Journey in Organizational Change Research and Practice. *Journal of Change Management*, 9(2), 127–142. <https://doi.org/10.1080/14697010902879079>
- Armenakis, A., Harris, S., Cole, M., Fillmer, L., & Self, D. (2007). A Top Management Team's Reactions to Organizational Transformation: The Diagnostic Benefits of Five Key Change Sentiments. *Journal of Change Management*, 7(3–4), 273–290. <https://doi.org/10.1080/14697010701771014>
- Armstrong, J. (1982). The value of formal planning for strategic decisions: review of empirical research. *Strategic Management Journal*, 3, 197–211.
- Ball, S., & Bowe, R. (1992). Subject departments and the 'implementation' of National Curriculum policy: an overview of the issues. *Journal of Curriculum Studies*, 24(2), 97–115. <https://doi.org/10.1080/0022027920240201>
- Barber, M. (2005). The virtue of accountability: System redesign, inspection, and incentives in the era of informed professionalism. *Journal of Education*, 185(1), 7–38. <https://doi.org/10.1177/002205740518500102>
- Baxter, J. A. (2013). Professional inspector or inspecting professional? Teachers as inspectors in a new regulatory regime for education in England. *Cambridge Journal of Education*, 43(4), 467–485. <https://doi.org/10.1080/0305764X.2013.819069>
- Behnke, K., & Steins, G. (2017). Principals' reactions to feedback received by school inspection: A longitudinal study. *Journal of Educational Change*, 18(1), 77–106.

<https://doi.org/10.1007/s10833-016-9275-7>

- Bengston, D., & Xu, Z. (1995). *Changing National Forest Values: a content analysis - Research Paper NC-323*. http://www.nrs.fs.fed.us/pubs/rp/rp_nc323.pdf
- Berry, F. S., & Wechsler, B. (1995). State agencies' experience with strategic planning: findings from a national survey. *Public Administration Review*, 55(2), 159. <https://doi.org/10.2307/977181>
- Bitan, K., Haep, A., & Steins, G. (2014). School inspections still in dispute – an exploratory study of school principals' perceptions of school inspections. *International Journal of Leadership in Education*, 18(4), 1–22. <https://doi.org/10.1080/13603124.2014.958199>
- Bloem, S. (2015). The OECD Directorate for Education as an independent knowledge producer through Pisa. In H. G. Kotthoff & E. Klerides (Eds.), *Governing Educational Spaces* (pp. 169–185). SensePublishers. https://doi.org/10.1007/978-94-6300-265-3_10
- Brier, A., & Hopp, B. (2011). Computer assisted text analysis in the social sciences. *Quality & Quantity*, 45(1), 103–128. <https://doi.org/10.1007/s11135-010-9350-8>
- Chabbott, C., & Elliott, E. J. (2003). Understanding others, educating ourselves: Getting more from international comparative studies in education. In *Social Sciences*. <https://doi.org/10.17226/10622>
- Chun, Y. H., & Rainey, H. G. (2005). Goal ambiguity and organizational performance in U.S. federal agencies. *Journal of Public Administration Research and Theory*, 15(4), 529–557. <https://doi.org/10.1093/jopart/mui030>
- Clarke, J., & Ozga, J. (2011). Governing by inspection? Comparing school inspection in Scotland and England. *Social Policy Association Conference*, 25.
- Coburn, C. (2001). Beyond decoupling: Rethinking the relationship between the institutional environment and the classroom. *Sociology of Education*, 77, 211–244. <https://doi.org/10.1177/003804070407700302>
- Coburn, C. (2005). Shaping teacher sensemaking: School leaders and the enactment of reading policy. *Educational Policy*, 19(3), 476–509. <https://doi.org/10.1177/0895904805276143>
- Cole, M. S., Harris, S., & Bernerth, J. B. (2006). Exploring the implications of vision, appropriateness, and execution of organizational change. *Leadership & Organization Development Journal*, 27(5), 352–367. <https://doi.org/10.1108/01437730610677963>
- Concurso de Supervisores Rio Negro, (2013).
- Resolución del Consejo Provincial de Educación de Río Negro N 1053, Pub. L. No. 1053 (1994).
- Conway, M. (2006). The Subjective Precision of Computers: A Methodological Comparison with Human Coding in Content Analysis. *Journalism & Mass Communication Quarterly*,

- 83(1), 186–200. <https://doi.org/10.1177/107769900608300112>
- Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Cuckle, P., Hodgson, J., & Broadhead, P. (1998). Investigating the relationship between OFSTED Inspections and school development planning. *School Leadership & Management, 18*(2), 271–283. <https://doi.org/10.1080/13632439869691>
- Darling-Hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Podolsky, A., & Stosich, E. L. (2016). *Pathways to new accountability through the Every Student Succeeds Act*. <http://learningpolicyinstitute.org/our-work/publications-resources/%0Apathways-new-accountability-every-student-succeeds-act>.
- De Vries, H., Elliott, M. N., Kanouse, D. E., & Teleki, S. S. (2008). Using Pooled Kappa to Summarize Interrater Agreement across Many Items. *Field Methods, 20*(3), 272–282. <https://doi.org/10.1177/1525822X08317166>
- de Wolf, I., & Janssens, F. (2007). Effects and side effects of inspections and accountability in education: An overview of empirical studies. *Oxford Review of Education, 33*(3), 379–396. <https://doi.org/10.1080/03054980701366207>
- Dedering, K., & Müller, S. (2011). School improvement through inspections? First empirical insights from Germany. *Journal of Educational Change, 12*(3), 301–322. <https://doi.org/10.1007/s10833-010-9151-9>
- Dedering, K., & Sowada, M. G. (2017). Reaching a conclusion—procedures and processes of judgement formation in school inspection teams. *Educational Assessment, Evaluation and Accountability, 29*(1), 5–22. <https://doi.org/10.1007/s11092-016-9246-9>
- Deng, Q., Hine, M., Ji, S., & Sur, S. (2019). Inside the black box of dictionary building for text analytics: a design science approach. *Journal of International Technology and Information Management, 27*(3), 119–159.
- Doud, J. (1995). Planning for school improvement: A curriculum model for school based evaluation. *Peabody Journal of Education, 70*, 175–187.
- Edgerton, A. K. (2019). The essence of ESSA: More control at the district level? *Phi Delta Kappan, 101*(2), 14–17. <https://doi.org/10.1177/0031721719879148>
- Education Inspectorate - Ministry of Education, C. and S. (2010). *Risk-based Inspection as of 2009 - Primary and Secondary Education*.
- Education Inspectorate - Ministry of Education, C. and S. (2017a). *Inspection framework primary education*.
- Education Inspectorate - Ministry of Education, C. and S. (2017b). *Inspection framework secondary education*.

- Ehren, M. (2016a). *Methods and modalities of effective school inspections* (M. Ehren (ed.)). Springer International Publishing. <https://doi.org/10.1007/978-3-319-31003-9>
- Ehren, M. (2016b). *Methods and Modalities of Effective School Inspections*. In M. C.M. Ehren (Ed.), *Methods and Modalities of Effective School Inspections*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-31003-9>
- Ehren, M., Altrichter, H., McNamara, G., & O'Hara, J. (2013). Impact of school inspections on improvement of schools—describing assumptions on causal mechanisms in six European countries. *Educ Asse Eval Acc*, 25, 3–43. <https://doi.org/10.1007/s11092-012-9156-4>
- Ehren, M., Gustafsson, J.-E., Altrichter, H., Skedsmo, G., Kemethofer, D., & Huber, S. (2015). Comparing effects and side effects of different school inspection systems across Europe. *Comparative Education*, 51(3), 375–400. <https://doi.org/10.1080/03050068.2015.1045769>
- Ehren, M., Perryman, J., & Shackleton, N. (2015a). School Effectiveness and School Improvement. *School Effectiveness and School Improvement - An International Journal of Research, Policy and Practice*, 26(2), 296–327.
- Ehren, M., Perryman, J., & Shackleton, N. (2015b). Setting expectations for good education: how Dutch school inspections drive improvement. *School Effectiveness and School Improvement*, 26(2), 296–327. <https://doi.org/10.1080/09243453.2014.936472>
- Ehren, M., & Shackleton, N. (2016). Risk-based school inspections: impact of targeted inspection approaches on Dutch secondary schools. *Educational Assessment, Evaluation and Accountability*, 28(4), 299–321. <https://doi.org/10.1007/s11092-016-9242-0>
- Ehren, M., & Visscher, A. (2006). TOWARDS A THEORY ON THE IMPACT OF SCHOOL INSPECTIONS. *British Journal of Educational Studies*, 54(1), 51–72. <https://doi.org/10.1111/j.1467-8527.2006.00333.x>
- Ehren, M., & Visscher, A. (2008). THE RELATIONSHIPS BETWEEN SCHOOL INSPECTIONS, SCHOOL CHARACTERISTICS AND SCHOOL IMPROVEMENT. *British Journal of Educational Studies*, 56(2), 205–227. <https://doi.org/10.1111/j.1467-8527.2008.00400.x>
- Fernandez, K. E. (2011). Evaluating school improvement plans and their affect on academic performance. *Educational Policy*, 25(2), 338–367. <https://doi.org/10.1177/0895904809351693>
- Figlio, D., & Loeb, S. (2011). School accountability. In *Handbook of the Economics of Education* (pp. 383–421).
- Fitchett, P., & Heafner, T. (2010). A national perspective on the effects of high-stakes testing and standardization on elementary social studies marginalization. *Theory & Research in Social Education*, 38(1), 114–130. <https://doi.org/10.1080/00933104.2010.10473418>
- Gagnon, D. J., & Schneider, J. (2019). Holistic school quality measurement and the future of

- accountability: Pilot-test results. *Educational Policy*, 33(5), 734–760.
<https://doi.org/10.1177/0895904817736631>
- Gilroy, P., & Wilcox, B. (1997). OFSTED, criteria and the nature of social understanding: A Wittgensteinian critique of the practice of educational judgement. *British Journal of Educational Studies*, 45(1), 22–38. <https://doi.org/10.1111/1467-8527.00034>
- Gioia, D., Thomas, J., Clark, S., & Chittipeddi, K. (1994). Symbolism and strategic change in academia: The dynamics of sensemaking and influence. *Organization Science*, 5(3), 363–383. <https://doi.org/10.1287/orsc.5.3.363>
- Glazerman, S. (2016). *The false dichotomy of school inspection*. Mathematica Policy Research - Blog Post. <https://www.mathematica-mpr.com/commentary/the-false-dichotomy-of-school-inspections>
- Gray, C., & Gardner, J. (1999). The impact of school inspections. *Oxford Review of Education*, 25(4), 455–468. <https://doi.org/10.1080/030549899103928>
- Gray, J., & Wilcox, B. (1995). In the aftermath of inspection: the nature and fate of inspection report recommendations. *Research Papers in Education*, 10(1), 1–18.
<https://doi.org/10.1080/0267152950100102>
- Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11(13), 255–274.
- Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267–297.
<https://doi.org/10.1093/pan/mps028>
- Grimolizzi-Jensen, C. J. (2018). Organizational change: Effect of motivational interviewing on readiness to change. *Journal of Change Management*, 18(1), 54–69.
<https://doi.org/10.1080/14697017.2017.1349162>
- Gustafsson, J.-E., Ehren, M., Conyngham, G., McNamara, G., Altrichter, H., & O'Hara, J. (2015). From inspection to quality: Ways in which school inspection influences change in schools. *Studies in Educational Evaluation*, 47, 47–57.
<https://doi.org/10.1016/j.stueduc.2015.07.002>
- Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. *Educational Administration, Policy, and Reform: Research and Measurement*, 3, 153–188.
- Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2), 297–327.
<https://doi.org/10.1002/pam.20091>
- Herscovitch, L., & Meyer, J. P. (2002). Commitment to organizational change: Extension of a

- three-component model. *Journal of Applied Psychology*, 87(3), 474–487.
<https://doi.org/10.1037/0021-9010.87.3.474>
- Hill, H. (2001). Policy is not enough: language and the interpretation of State standards. *American Educational Research Journal*, 38(2), 289–318.
<https://doi.org/10.3102/00028312038002289>
- Hines, R. T. (2017). *An Exploration of the Effects of School Improvement Planning and Feedback Systems: School Performance in North Carolina*.
- Holt, D., Armenakis, A., Feild, H., & Harris, S. (2007). Readiness for Organizational Change. *The Journal of Applied Behavioral Science*, 43(2), 232–255.
<https://doi.org/10.1177/0021886306295295>
- Husfeldt, V. (2011). Wirkungen und wirksamkeit der externen schulevaluation; uberblick zum stand der forschung [The impact of school inspection - Does it really work? State of research]. *Zeitschrift Für Erziehungswissenschaft*, 14(2), 259–282.
<https://doi.org/10.1007/s11618-011-0204-5>
- Hussain, I. (2015). Subjective Performance Evaluation in the Public Sector Evidence from School Inspections. *The Journal of Human Resources*, 50(1), 189–221.
- Jacob, B. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5–6), 761–796.
<https://doi.org/10.1016/j.jpubeco.2004.08.004>
- Jones, K., & Tymms, P. (2014). Ofsted’s role in promoting school improvement: the mechanisms of the school inspection system in England. *Oxford Review of Education*, 40(3), 315–330.
- Jones, K., Tymms, P., Kemethofer, D., O’Hara, J., McNamara, G., Huber, S., Myrberg, E., Skedsmo, G., & Greger, D. (2017). The unintended consequences of school inspection: the prevalence of inspection side-effects in Austria, the Czech Republic, England, Ireland, the Netherlands, Sweden, and Switzerland. *Oxford Review of Education*, 43(6), 805–822.
<https://doi.org/10.1080/03054985.2017.1352499>
- Kaplan, S., & Orlikowski, W. J. (2013). Temporal Work in Strategy Making. *Organization Science*, 24(4), 965–995. <https://doi.org/10.1287/orsc.1120.0792>
- Klein, A. (2016). *School inspections offer a diagnostic look at quality*. Education Week.
<https://www.edweek.org/ew/articles/2016/09/28/school-inspections-offer-a-diagnostic-look-at.html>
- Klerks, M. (2012). *The effect of school inspections: a systematic review*. <http://janbri.nl/wp-content/uploads/2014/12/ORD-paper-2012-Review-Effect-School-Inspections-MKLERKS.pdf>
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A

- historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>
- Koretz, D. (2008). *Measuring up*. Harvard University Press.
- Krippendorff, K. (2013). *Content analysis: An introduction to Its methodology* (3rd ed.). SAGE Publications.
- Ladd, H. F. (2016). *Now is the time to experiment with inspections for school accountability*. Brookings. <https://www.brookings.edu/blog/brown-center-chalkboard/2016/05/26/now-is-the-time-to-experiment-with-inspections-for-school-accountability/>
- Ladd, H. F. (2017). NCLB: RESPONSE TO JACOB. *Journal of Policy Analysis and Management*, 36(2), 477–480. <https://doi.org/10.1002/pam.21979>
- Ladd, H. F., & Figlio, D. (2008). School accountability and student achievement. In *Handbook of research in education finance and policy* (pp. 166–182).
- Lee, J., & Fitz, J. (1997). HMI and OFSTED: evolution or revolution in school inspection. *British Journal of Educational Studies*, 45(1), 39–52. <https://doi.org/10.1111/1467-8527.00035>
- Lewin, A. Y., & Minton, J. W. (1986). Determining Organizational Effectiveness: Another Look, and an Agenda for Research. *Management Science*, 32(5), 514–538. <https://doi.org/10.1287/mnsc.32.5.514>
- Lindgren, J. (2015). The front and back stages of swedish school inspection: opening the black box of judgment. *Scandinavian Journal of Educational Research*, 59(1), 58–76. <https://doi.org/10.1080/00313831.2013.838803>
- Luginbuhl, R., Webbink, D., & de Wolf, I. (2009). Do Inspections Improve Primary School Performance? *Analysis*, 31(3), 221–237. <https://doi.org/10.3102/0162373709338315>
- Maitlis, S. (2005). The Social Processes of Organizational Sensemaking. *The Academy of Management Journal*, 48(1), 21–49. <https://doi.org/10.2307/20159639>
- Maitlis, S., & Christianson, M. (2014). Sensemaking in organizations: Taking stocks and moving forward. *The Academy of Management Annals*, 8(1), 57–125. <https://doi.org/10.1080/19416520.2014.873177>
- March, J. G., & Olsen, J. P. (2011). The Logic of Appropriateness. In R. E. Goodin (Ed.), *The Oxford Handbook of Political Science* (pp. 1–22). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199604456.013.0024>
- Mathis, W., & Trujillo, T. (2016). *Lessons from NCLB for the Every Student Succeeds Act*. <http://nepc.colorado.edu/%0Apublication/lessons-from-NCLB>
- Matthews, P., & Sammons, P. (2004). *Improvement through inspection. An evaluation of the*

impact of Ofsted's work. Ofsted.

- Matthews, Peter, Holmes, J. R., Vickers, P., & Corporaal, B. (1998). Aspects of the reliability and validity of school inspection judgements of teaching quality. *Educational Research and Evaluation, 4*(2), 167–188. <https://doi.org/10.1076/edre.4.2.167.6959>
- McDonnell, L. (2008). The politics of educational accountability: Can the clock be turned back? In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability*. Routledge.
- McDonnell, L. (2013). Educational accountability and policy feedback. *Educational Policy, 27*(2), 170–189. <https://doi.org/10.1177/0895904812465119>
- Meyers, C. V., & VanGronigen, B. A. (2019). A lack of authentic school improvement plan development. *Journal of Educational Administration, 57*(3), 261–278. <https://doi.org/10.1108/JEA-09-2018-0154>
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3th ed.). SAGE Publications.
- Millett, A., & Johnson, D. C. (1998). Expertise or “baggage”? What helps inspectors to inspect primary mathematics? *British Educational Research Journal, 24*(5), 503–518. <https://doi.org/10.1080/0141192980240502>
- Mintrop, H., MacLellan, A. M., & Quintero, M. F. (2001). School improvement plans in schools on probation: A comparative content analysis across three accountability systems. *Educational Administration Quarterly, 37*(2), 197–218. <https://doi.org/10.1177/00131610121969299>
- Morse, J. (2010). Procedures and practice of mixed method design - Maintaining control, rigor, and complexity. In A. M. Tashakkori & C. B. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 339–352). SAGE Publications.
- Neuendorf, K. A. (2017). *The content analysis guidebook*. SAGE Publications, Inc. <https://doi.org/10.4135/9781071802878>
- Nusche, D., Braun, H., Halász, G., & Santiago, P. (2014). *OECD Reviews of Evaluation and Assessment in Education: Netherlands 2014*. OECD. <https://doi.org/10.1787/9789264211940-en>
- OECD, O. for E. C. and D. (2015). *Education at a glance 2015 - OECD Indicators*. <https://doi.org/10.1787/19991487>
- Ouston, J., Fidler, B., & Earley, P. (1997). What do schools do after OFSTED school inspections-or before? *School Leadership & Management, 17*(1), 95–104. <https://doi.org/10.1080/13632439770195>
- Penninckx, M., & Vanhoof, J. (2015). Insights gained by schools and emotional consequences of

- school inspections. A review of evidence. *School Leadership & Management*, 35(5), 477–501. <https://doi.org/10.1080/13632434.2015.1107036>
- Penninckx, M., Vanhoof, J., De Maeyer, S., & Van Petegem, P. (2014). Exploring and explaining the effects of being inspected. *Educational Studies*, 40(4), 456–472. <https://doi.org/10.1080/03055698.2014.930343>
- Penninckx, M., Vanhoof, J., De Maeyer, S., & Van Petegem, P. (2015). Effects and side effects of Flemish school inspection. *Educational Management Administration & Leadership*. <https://doi.org/10.1177/1741143215570305>
- Perryman, J. (2007). Inspection and emotion. *Cambridge Journal of Education*, 37(2), 173–190. <https://doi.org/10.1080/03057640701372418>
- Perryman, J. (2009). Inspection and the fabrication of professional and performative processes. *Journal of Education Policy*, 24(5), 611–631.
- Phillips, D., & Schweisfurth, M. (2014). *Comparative and international education - An introduction to theory, methods , and practice* (2nd ed.). Group, Continuum International Publishing.
- Piderit, S. K. (2000). Rethinking resistance and recognizing ambivalence: A multidimensional view of attitudes toward an organizational change. *The Academy of Management Review*, 25(4), 783. <https://doi.org/10.2307/259206>
- Pond, S., Armenakis, A., & Green, S. (1984). The Importance of Employee Expectations in Organizational Diagnosis. *The Journal of Applied Behavioral Science*, 20(2), 167–180. <https://doi.org/10.1177/002188638402000207>
- Porac, J. F., Thomas, H., & Baden-Fuller, C. (1989). COMPETITIVE GROUPS AS COGNITIVE COMMUNITIES: THE CASE OF SCOTTISH KNITWEAR MANUFACTURERS. *Journal of Management Studies*, 26(4), 397–416. <https://doi.org/10.1111/j.1467-6486.1989.tb00736.x>
- Portz, J., & Beauchamp, N. (2020). Educational Accountability and State ESSA Plans. *Educational Policy*, 089590482091736. <https://doi.org/10.1177/0895904820917364>
- Ravitch, D. (2016). *The death and life of the great American school system: How testing and choice are undermining education*. Basic Books.
- Redding, C., & Searby, L. (2020). The Map Is Not the Territory: Considering the Role of School Improvement Plans in Turnaround Schools. *Journal of Cases in Educational Leadership*, 23(3), 63–75. <https://doi.org/10.1177/1555458920938854>
- Riffe, D., Lacy, S., & Fico, F. (2014). *Analyzing media messages: Using quantitative content analysis in research*. Routledge.
- Rigby, J. G. (2015). Principals’ sensemaking and enactment of teacher evaluation. *Journal of*

- Educational Administration*, 53(3), 374–392. <https://doi.org/10.1108/JEA-04-2014-0051>
- Rosenthal, L. (2004). Do school inspections improve school quality? Ofsted inspections and school examination results in the UK. *Economics of Education Review*, 23, 143–151.
- Rothstein, R., Jacobsen, R., & Wilder, T. (2008). *Grading education: Getting accountability right*. Economic Policy Institute and Teacher College Press.
- Rouleau, L. (2005). Micro-practices of strategic sensemaking and sensegiving: How middle managers interpret and sell change every day. *Journal of Management Studies*, 42(7), 1413–1441.
- Rutz, S., Mathew, D., Robben, P., & Bont, A. (2017). Enhancing responsiveness and consistency: Comparing the collective use of discretion and discretionary room at inspectorates in England and the Netherlands. *Regulation & Governance*, 11(1), 81–94. <https://doi.org/10.1111/regg.12101>
- Ryan, K., Gandha, T., & Ahn, J. (2013). School self-evaluation and inspection for improving U.S. schools? In *National Education Policy Center*. <http://nepc.colorado.edu/publication/school-self-evaluation>
- Sandberg, J., & Tsoukas, H. (2015). Making sense of the sensemaking perspective: Its constituents, limitations, and opportunities for further development. *Journal of Organizational Behavior*, 36(S1), S6–S32. <https://doi.org/10.1002/job.1937>
- Scheerens, J., Ehren, M., Slegers, P., & Leeuw, R. de. (2012). *OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes*.
- Shaw, I., Newton, D. P., Aitkin, M., & Darnell, R. (2003). Do OFSTED Inspections of Secondary Schools Make a Difference to GCSE Results? *British Educational Research Journal*, 29(1), 63–75.
- Spillane, J. P. (1999). External reform initiatives and teachers' efforts to reconstruct their practice: The mediating role of teachers' zones of enactment. *Journal of Curriculum Studies*, 31(2), 1–33. <https://doi.org/10.1080/002202799183205>
- Spillane, J. P., Parise, L. M., & Sherer, J. Z. (2011). Organizational Routines as Coupling Mechanisms. *American Educational Research Journal*, 48(3), 586–619. <https://doi.org/10.3102/0002831210385102>
- Spillane, J. P., Reiser, 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 education policy implementation* (pp. 47–64). State University of New York Press, Albany.
- Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of Educational Research*, 72(3), 387–341. <https://doi.org/10.3102/00346543072003387>

- Stiglitz, J. (2000). *Economics of the public sector* (3rd ed.). Norton.
- Strunk, K. O., Marsh, J. A., Bush-Mecenas, S., & Duque, M. R. (2016). The Best Laid Plans. *Educational Administration Quarterly*, *52*(2), 259–309. <https://doi.org/10.1177/0013161X15616864>
- Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating qualitative and quantitative approaches in the social and behavioral sciences*. SAGE.
- Teddlie, C., & Yu, F. (2007). Mixed methods sampling : A typology with examples. *Journal of Mixed Methods Research*, *1*(1), 77–100. <https://doi.org/10.1177/1558689806292430>
- UNESCO. (2017). *Global Education Monitoring Report - Accountability in education: Meeting our commitments*.
- van Bruggen, J. C. (2010). *Inspectorates of Education in Europe; some comparative remarks about their tasks and work*.
- van der Sluis, M. E., Reezigt, G. J., & Borghans, L. (2017). Implementing New Public Management in Educational Policy. *Educational Policy*, *31*(3), 303–329.
- Vavrus, F. K., & Bartlett, L. (2016). *Rethinking case study research: A comparative approach* (1st ed.). Routledge.
- Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2010). Using school performance feedback: perceptions of primary school principals. *School Effectiveness and School Improvement*, *21*(2), 167–188. <https://doi.org/10.1080/09243450903396005>
- Visscher, A. J., & Coe, R. (2003). School performance feedback Systems: conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement*, *14*(3), 321–349. <https://doi.org/10.1076/sesi.14.3.321.15842>
- Weick, K. E. (1995). *Sensemaking in organizations*. SAGE Publications.
- Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, *16*(4), 409–421. <https://doi.org/10.1287/orsc.1050.0133>
- Weiner, B. J. (2009). A theory of organizational readiness for change. *Implementation Science*, *4*(1), 67. <https://doi.org/10.1186/1748-5908-4-67>
- Woods, P., & Jeffrey, B. (1998). Choosing Positions: Living the Contradictions of OFSTED. *British Journal of Sociology of Education*, *19*(4), 547–570. <https://doi.org/10.1080/0142569980190406>

Paper 2: Principals' Attitudes towards School Inspection in a U.S. District: Contribution to Sustained School Reform

Introduction

Accountability aims to improve education quality by providing oversight of school performance. High-stakes testing and school inspection (SI) are the most widely used accountability approaches around the world. In the United States, high-stakes testing is the primary instrument for accountability, although a few states and cities have experimented with SI. Both approaches aim to influence stakeholders' actions to spur improved outcomes. Yet, these approaches have different theories of action for improvement. In the case of inspection, principals acceptance of feedback is an intermediate step before implementing reforms (Ehren et al., 2013). Principals' attitudes towards inspection feedback likely influences their readiness for taking improvement actions. Further, there are key attitudes from those who implement reforms in an organization that are critical to its sustainability (Armenakis & Harris, 2009). This study assesses how inspections affect principals' attitudes associated with lasting change.

Many countries use SI as the main mechanism for accountability. Rather than relying solely on standardized tests, inspection considers a variety of evidence on school processes and practices. These evaluations are conducted by inspectors, which in the best cases are experienced educators who maintain close contact with schools. Inspectors are able to provide informed feedback taking into consideration the local context, system-wide evidence, and academic research (Altrichter & Kemethofer, 2015; UNESCO, 2017). By visiting schools and observing practices, inspectors can obtain a deeper understanding of the existing capacity for improvement and existing efforts to improve performance. Such observations, in combination with test scores, enable a more in-depth

understanding of school operations and nuanced insights into what might help or hinder improvement (Barber, 2005).

School leaders' perceptions about the validity of inspection influence their response to results (e.g. Bitan et al., 2014). If inspections are viewed as relevant, accurate, and fair, then school leaders might be compelled to implement suggestions. Despite potential advantages of inspection, few studies investigate its role in supporting school reform. It is not clear whether it is an effective mechanism for long-term changes. No peer-reviewed study provides empirical evidence of the ability of inspection to enable sustained institutional change.

Bridging insights from the organizational change literature in the education field, this paper assesses how inspection affects the attitudes and sentiments of school principals (agents of change), whose decisions are crucial for achieving lasting institutional change (Armenakis & Harris, 2009). These attitudes include 1) beliefs about the diagnosis effectiveness, 2) perceived appropriateness of feedback, and 3) readiness for making changes (Armenakis & Harris, 2009; Holt et al., 2007; Weiner, 2009). I draw on case study data collected in a large urban district that is one of the few districts in the U.S. to use school inspection. With one of the most well-developed inspection systems in the country, this district presents a unique research opportunity to understand how school leaders perceive and act upon inspection feedback. Test-based accountability has been subject to debate and criticism. It is unclear if inspection is viewed by school administrators as more accurate and fair.

This study examines how inspection influences school principals' attitudes that are associated with lasting institutional change. Understanding underlying attitudes can shed light on whether inspection offers potential to create long-term school improvements. I address the following research questions:

- 1) How do inspection and test-based accountability compare in terms of principals' perceived accuracy in evaluating areas for improvement?

- 2) Do school inspections promote principals' attitudes associated with lasting institutional change?
 - a. Do principals feel that inspections provide an effective diagnosis of the main challenges and factors that hinder improvement?
 - b. Do principals perceive suggested reforms as appropriate?
 - c. Do inspections promote principals' readiness for change?

Background

As the implementation of Every Student Success Act in the U.S. has sparked debate over the design of accountability schemes, inspection could be viewed as an alternative option for test-based accountability. Yet, one major barrier to expanded inspection in the U.S. is limited empirical evidence about its impact (Husfeldt, 2011; Klerks, 2012). Furthermore, there is no evidence about its role in supporting lasting school reforms.

Test-based accountability and inspection systems rely on different sets of assumptions regarding school stakeholder behavior. In addition, these two systems of accountability may influence school stakeholders' attitudes and decisions in different ways. Theories of action behind these contrasting accountability schemes shed light on these differences.

The theory of action of test-based accountability for improvement is grounded in the principal-agent problem. This problem theorizes that inefficiencies in the public sector arise when *agents* (i.e. government bureaucrats, such as school leaders) focus on their own interests, which are not necessarily the same as those of the *principals* (i.e. the citizens they are supposed to serve) (Stiglitz, 2000). This theory reasons that a correct incentive structure is an effective way of aligning the interests of *agents* with those of the *principals*. Grounded in this rationale, test-based accountability uses standardized test results as a metric for determining rewards and/or sanctions to schools based on the aggregate results in these tests. The set of incentives can be explicit actions enforced by the government (e.g. positive incentives such as bonuses; negative incentives such as

school closure) or implicit through community pressure to improve after test results are published (Figlio & Loeb, 2011). Thus, standardized test results can provide an incentive structure for schools to make a concerted effort to improve the outcomes that are measured (Jacob, 2005; Ladd & Figlio, 2008).

Despite its promise, the rationale of test-based accountability oftentimes is inadequate when there are local circumstances that prevent administrators and teachers from responding to incentives as policymakers intended. Barriers can include low-resource settings or cultural and language factors (Figlio & Loeb, 2011; Koretz, 2008). Also, pressure to improve test results in the short run may compel school administrators to implement changes that improve measured outcomes. Yet, these short-term reforms may displace more holistic strategies that require a longer timeframe to produce results (e.g. Figlio & Loeb, 2011; Ladd & Figlio, 2008). As a result, administrators of low performing schools might feel pressure to focus on narrow, short-term improvements and implement changes that they do not fully support (Ladd & Figlio, 2008). This might reduce the sustainability and long-term success of school reforms.

In contrast, SI relies on different mechanisms for school improvement. There is less consensus on theories of action for SI, partly due to the wide variety of inspection systems (Husfeldt, 2011). Ehren et al (2013) developed a comprehensive framework, which describes world wide variations in school inspection.. According to this framework, accountability pressure plays a role for school improvement, especially high-stakes inspection systems. Feedback is another mechanism for school improvement for all types of inspection arrangements. This is the focus of this study. In theory, feedback compels schools to make improvements in areas evaluated as weak. For this theory of action to function, it is crucial for school inspectors to provide relevant input for school improvement and for school stakeholders accept the feedback and to act on it.

Unlike test-based accountability, on-site feedback mechanisms have the potential to offer relevant information and insights as a basis for practitioner-led improvement actions (Visscher &

Coe, 2003). Information is more localized and also considers the specific context, processes, and practices (Altrichter & Kemethofer, 2015). This entails a relative advantage for its potential to cover a much broader range of education quality aspects, to account for the multi-dimensional nature of school quality (Ryan, Gandha, & Ahn, 2013). Thus, inspections can offer nuanced insight into what might hinder or enable school improvement (Barber, 2005). While test results show that performance is poor, inspection information has greater ability to determine why. This can provide schools with greater ability to understand specifics about problems areas and to take improvement actions based on a more holistic evaluation.

In addition, inspection processes have the potential for opening dialogue between the district leaders and school principals. This interaction can potentially shed light on specific school problems and enable inspectors to propose alternative reforms that allow administrators to retain ongoing successful practices. This in turn can lead to smoother incorporation of reforms and avoid abrupt changes in administration and conflicts with school culture (Ehren, 2016).

As the literature above describes, school inspections can provide relevant information and insights that lead to school improvements. However, the success of this process hinges on principals accepting and acting upon inspection feedback. Extant research has not yet evaluated whether and how school principals evaluate and use inspection feedback to implement long-term reforms.

Literature Review

Few studies address the influence of principal attitudes toward inspection on institutional change in schools (Behnke & Steins, 2017; Bitan et al., 2014; C. Gray & Gardner, 1999). Other studies mention attitudinal reactions toward inspections (e.g. de Wolf & Janssens, 2007; Jones et al., 2017; Penninckx & Vanhoof, 2015). Generalizing findings from this literature is challenging due to studies addressing a wide variety of attitudes and side-effects in countries that have varied inspection systems.

Principals' attitudes towards inspection influence their acceptance of feedback. Most prior studies find positive sentiments towards inspection feedback (Behnke & Steins, 2017; Bitan et al., 2014; C. Gray & Gardner, 1999). Principals feel inspection feedback is relevant (Bitan et al., 2014), accurate and fair (C. Gray & Gardner, 1999), constructive, and supportive (Behnke & Steins, 2017). In some cases, principals positive attitudes are due to inspection serving as a source of legitimacy for reform they propose (Behnke & Steins, 2017; Ehren, Perryman, et al., 2015a). Negative sentiments have been found in literature focused on effects and side-effects of inspection. These sentiments include emotional distress (i.e. stress, anxiety, fear), concerns due to work overload, tension between staff members, and reduced self-efficacy (e.g. de Wolf & Janssens, 2007; Jones et al., 2017; Penninckx et al., 2015; Penninckx & Vanhoof, 2015; Perryman, 2007, 2009). Much of the past literature addressing the emotional consequences of inspections has been conducted in the UK (Penninckx & Vanhoof, 2015), which has a high-stakes inspection system. No study was found in this review indicating an overall negative attitude from school principals towards inspection.

The way in which feedback is delivered can influence principal attitudes. Findings indicate that administrators tend to be more ready to take improvement actions when inspection identifies weaknesses and creates an ongoing dialogue between inspectors and school administrators (Bitan et al., 2014; Ehren & Visscher, 2008). Past work has also consistently shown that inspection feedback better enables school development and improvement when it is combined with agreement on targets with schools (Behnke & Steins, 2017; Dederling & Müller, 2011; Ehren, Perryman, et al., 2015b; Ehren & Visscher, 2008). School leaders will more readily implement improvement plans if inspectors report weaknesses in a straightforward manner, make written recommendations, and encourage the development of improvement plans (Ehren & Visscher, 2006, 2008; Penninckx et al., 2014). Yet, critical feedback can have negative effects on the self-worth of those evaluated that might lead to dismissing the feedback despite its value (Kluger & DeNisi, 1996).

Accountability pressures also may influence principals' attitudes towards inspection feedback. Higher accountability pressure can lead to greater sensitivity in the schools regarding inspectors' quality expectations, which are translated in developmental activities (Altrichter & Kemethofer, 2015). Nonetheless, higher pressure is also associated with unintended consequences such as narrowing the curriculum and instructional strategies (Altrichter & Kemethofer, 2015; Jones et al., 2017).

These studies have explored a wide range of factors that demonstrated an association between inspection feedback with school stakeholders' attitudes and undertaking improvement action. Yet, none have focused on the specific attitudes that are relevant lasting institutional change. This study addresses this gap in the literature.

Theoretical Framework

The organizational change literature offers useful constructs to analyze the role that attitudes and beliefs in promoting change in schools. Over time, this literature has empirically identified key factors that facilitate sustained organizational change. Especially relevant is a paper by Armenakis and Harris (2009), which reviews studies from the past 30 years. Building on their work, I focus on factors relevant for educational systems. Within this framework, I consider the educational system as an organization, with school principals as local agents of change.

Past literature has identified beliefs or sentiments that underlie actors motives to support change within organizations and increase the likelihood of sustained change (e.g. Armenakis et al., 2007; Lewin & Minton, 1986). Reforms are more likely to be successful when actors believe that changes are needed to improve organizational performance. In addition, specific changes gain support when actors trust that they are suitable to address problems and can be successfully implemented. Commitment to reform is another critical factor associated with sustained change. In an influential paper, Herscovitch and Meyer (2002) argue that what motivates change is associated with the level of commitment. They propose that in an organization, members' commitment to

change occurs because they “want to” (*affective commitment*), “have to” in order to avoid failure (*continuance commitment*), or “ought to” out of a sense of obligation (*normative commitment*).

Affective and continuance commitment are found to be associated with a higher level of commitment (Herscovitch & Meyer(2002).

Building on these findings, the organizational change literature has identified central factors for sustained organizational change:

- 1) Diagnosis effectiveness. An effective organizational diagnosis is accurate, recognizes the main problem and identifies the root cause (Armenakis & Harris, 2009). This is a critical condition to convince an individual that a change is needed (Armenakis & Harris, 2009; Pond et al., 1984).
- 2) Perceived appropriateness of feedback: An effective organizational diagnosis leads to changes that are appropriate to address the problems in an organization (Armenakis & Harris, 2009; Holt et al., 2007). The appropriateness capture how course of actions are perceived as natural, rightful, or legitimate to pursue the organizational need for change and vision (Cole et al., 2006; March & Olsen, 2011). It is possible that a vision is embraced by individuals, but they do not believe that a specific change leads to fulfill that vision (Cole et al., 2006).
- 3) Readiness for change. Reforms must be perceived to be appropriate to succeed (Holt et al., 2007). Yet, organizational members can have ambivalent attitudes. Conflicts can arise across multiple dimensions – cognitive, emotional, and behavioral (Grimolizzi-Jensen, 2018; Piderit, 2000). If there is insufficient planning to prevent stalled decisions, reforms are less likely to succeed (Armenakis & Harris, 2009). Readiness of change refers to the commitment and belief of efficacy towards organizational changes (Weiner, 2009). This belief is necessary to overcome inertia due to ambivalent attitudes and increase the chances of sustainable organizational change.

These constructs and findings directly inform the research questions. I examine principals' perceptions of diagnosis effectiveness, both for inspection and the district accountability framework, which is largely based on standardized tests. Then, I assess principals' sentiments regarding the appropriateness of the SI feedback. Finally, I assess whether the diagnosis and changes based on the feedback influence principals' readiness for change. These are key conditions associated with sustainable institutional change.

Case Study: Description of School Inspection System

Similar to all U.S. districts, the selected case study has a high-stakes accountability system, which relies heavily on standardized testing. All schools are part of the district *accountability framework*. This framework rates schools based on a series of performance metrics, where standardized test results are the most influential component. In addition, the district also features support mechanisms, including principals' supervisors and instructional partners. For almost a decade, the district has combined their accountability framework with school inspection. Low-performing schools are targeted for inspection visits, yet other schools can opt-in. Since 2012, the inspections have been led by a private company that holds a contract with the district.

Inspections are conducted for accountability and support purposes. Qualitative evidence is compiled via a school visit, which is conducted by groups of three to four inspectors. Inspectors include personnel from an outside contractor and certified inspectors from the District Department of Education. The onsite inspections last one or two days and include document reviews, classroom observations, focus groups, and interviews with school staff, students, and parents. Inspection areas include instruction, students' opportunities to learn, educators' opportunities to learn, and leadership and community. Evaluation criteria emphasize instructional quality, support, and assessment, classroom climate, and school-wide practices and culture.

To conclude the school visit, a feedback meeting between inspectors and school administrators allows discussion of findings and improvement strategies. A written report summarizing conclusions is provided to schools, which includes suggestions for priority areas.

Methods

This study aims to evaluate whether SI influence a series of key factors important for lasting institutional change. It uses a large U.S. district as a case study to investigate how inspection influences principal attitudes towards sustainable institutional change. In-depth semi-structured interviews were conducted with 20 principals. All principals from schools in the tier-support system were invited to participate. Schools within the district’s tier-support system are primarily low-performing schools. Most have had at least one inspection visit (55 out of 78 schools in the tier-support system). Interviews were completed with 16 principals from inspected schools and 4 from not yet inspected schools. Participants were informed that interview responses were anonymous, and a pseudonym would be used to cite them. Principals were given a US\$ 25 gift card after participation. Descriptive data about interviewees are presented in Table 3.

Table 3. Principals’ Experience and Education

	Inspected schools		Not inspected schools	
	Number	(%)	Number	(%)
Principal experience				
0 to 4	4	(25.0)	4	(100.0)
5 to 8	7	(43.8)	-	-
9 to 12	3	(18.8)	-	-
13+	2	(12.5)	-	-
Years working in the school				
0 to 4	9	(56.3)	4	(100.0)
5 to 8	4	(25.0)	-	-
9 to 12	3	(18.8)	-	-
13+	-		-	-

Table 3 (cont'd)

	Inspected schools		Not inspected schools	
	Number (%)		Number (%)	
Teacher experience				
0 to 4	3	(18.8)	1	(25.0)
5 to 8	9	(56.3)	3	(25.0)
9 to 12	3	(18.8)	-	-
13+	1	(6.3)	-	-
Highest Degree				
Masters	12	(75.0)	4	(100.0)
Specialist	2	(12.5)	-	-
Ph.D. / Doctorate	2	(18.8)	-	-
n	16		4	

Interviews inquire about principals' attitudes towards inspection and its influence on institutional change. Inspection is the main focus, yet some questions address the district accountability framework, which is used in all schools. The district accountability framework relies heavily on high-stakes test results to rate schools. This allows for a direct comparison of principals' views on these two accountability approaches. Interviews inquired about what the principals learned from the SIs, the perceived legitimacy of feedback, and how the process aligned with the principal's vision for the school, specific programing, and long-term goals. Principals discussed their motivations for change, reforms already implemented based on the inspections, and their expectations for reform success and continuity. The interview protocol is presented in Appendix A.

Interview questions were informed by the organizational change literature (see the Theoretical Framework section). Interviews were coded primarily through deductive coding, although inductive coding was used to a lesser extent. Deductive codes were formulated using an *a priori* scheme developed based on constructs from organizational change theory (e.g. Armenakis,

Bernerth, et al., 2007; Holt et al., 2007). Constructs were adapted to the context of school reform implementation. Principals act both as agents of change and as change recipients. Changes stemming from the SI are mostly decided by the principals—as opposed to top-down changes typically considered in the organizational change literature. Inductive codes were added to reflect key ideas being investigated that were not captured by *a priori* codes. These codes were developed to analyze principals' perceptions of SI feedback, for example, its usefulness and how it influenced reform implementation in the school. Details of the coding scheme are provided in Appendix B.

To ensure code reliability, an independent-coder method was employed. First, a subset of interview transcripts was independently coded by two researchers using DEDOOSE software. Then, codes were compared for agreement. An iterative process was followed until at least 80% agreement was achieved on the pooled Kohen Kappa (De Vries et al., 2008). After this reliability check, the remaining transcripts were coded by the lead author. Using DEDOOSE, patterns were identified both within and across interviews. Data were examined visually in the form of cross-tabulations and frequency charts in order to determine the presence and absence of codes both within and across interviews.

A limitation of this study is related to relying on self-reporting of planned reforms. The interviews rely on self-reported reforms and therefore are limited to actions that principals can recall or decide to report. The study does not capture implemented reforms due to SI. Another possible limitation is that principals' attitudes towards inspection might be positively biased, to satisfy the district. Several factors alleviate all these concerns. Principals were informed that interviews were anonymous and that no identifying information for the school would be reported. There are indications that principals' responses were quite candid and honest. Many respondents did not shy away from expressing critical views of the district, its accountability system, and school inspections. I had no indication in the interviews that principals were using coded language or seemed hesitant to respond to particular questions. This alleviates some concern regarding self-censoring. In addition,

inspection visits were relatively recent and a major component of their planning process; principals provided detail about their experiences, which lessens concern over recall bias.

Results

Results indicate that principals hold attitudes towards inspection that are associated with lasting institutional change. Most principals from inspected schools exhibit positive attitudes and believe that on-site visits enable inspectors to attain a holistic understanding. In contrast, the district accountability framework alone has limited ability to understand the schools and perform an accurate assessment, as acknowledged by a majority of principals.

Interviewees believe that the inspection diagnosis is effective (69%), the feedback is appropriate for their schools (69%), and they are ready to support changes based on the inspections (75%). These attitudes that were strongly interrelated within cases. A minority of principals (25%) hold negative attitudes towards inspection and question its validity. These negative attitudes are driven by perceptions that changes suggested by inspections are not well-aligned with school goals and have a negative tone. They are not ready to make changes based on inspection feedback. Results are summarized in Table 4.

Table 4. Principals’ Attitudes towards School Inspection

Principal	Diagnosis Effectiveness	Feedback Appropriateness	Readiness for changes
Nicholas	yes	yes	yes
Mary	yes	mostly	yes
Tyler	yes	mostly	yes
Thomas	yes	mostly	yes
Monica	mostly	yes	yes
Linda	mostly	yes	yes
Sarah	mostly	yes	yes
Mark	yes	yes	mostly
Amy	yes	mostly	mostly

Table 4 (cont'd)

Principal	Diagnosis Effectiveness	Feedback Appropriateness	Readiness for changes
Heather	no mention	yes	mostly
Matthew	mostly	mostly	mostly
David	mostly	mostly not	mostly
Sebastian	mostly not	no	mostly not
Brian	no	no	no
Paul	no mention	no	no
Ashley	no mention	no	no
Yes	69%	69%	75%
No	19%	31%	25%

Perceptions about the District Diagnosis Effectiveness excluding Inspection

Besides inspection, all schools are part of the district *accountability framework*. This framework rate schools based on a series of performance metrics, where standardized test results are the most influential component. In addition, the principals’ supervisors also assess schools to provide support and keep them accountable.

Half of the principals think that this framework can produce an accurate evaluation. Other principals argue that this information is limited, and assessments are unlikely to be accurate. The consensus among principals is that the district relies heavily on quantitative metrics from the district accountability framework for diagnostic purposes, yet these data might not best characterize specific conditions within schools. Most principals feel that accurate assessment could be made by the few people that know the school. Nearly 75% of principals acknowledged that several individuals are deeply familiar with their school, including principals’ supervisors and, to a lesser degree, instructional leaders and senior district leaders.

Some interviewees point out the downside of information not being widely available and residing in “pockets” within the district. Principal Thomas voices this concern, explaining that an

diagnosis of the school “would probably not be very accurate because the information lives in pockets instead of it living in some centralized place, where I have confidence that everybody has access to the information that they need to have.” Furthermore, several principals note that the accessibility and continuity of this concentrated knowledge is vulnerable to high turnover in school leadership.

Limitations of district knowledge of school conditions are acknowledged by nearly all principals interviewed (19 out of 20). When asked specifically about what the district overlooks with the performance framework, principals most frequently mentioned that relying on test scores can fail to capture specific school challenges, strengths, and ongoing initiatives. More than half of principals raised this view. For example, some principals described limitations of district metrics to capture impediments to school improvement. Often, assessments are unable to identify specific factors that limit student achievement. Several principals attribute this to the district not having full knowledge of the student body and communities, particularly in terms of socio-emotional factors. Principals Mary and Linda illustrate these perspectives:

Principal Mary: [The central office] looks at the data. ... They have plenty of data. ... But if you want to come in and dig deep into why kids don't come to school, [and] why kids are not succeeding ... then I think they may not be as familiar.... Other social and emotional issues that factor into the students' success, I would say does not come out very well in a data chart.

Principal Linda: I'm not sure how they're looking at the data because when you have children who come to this school with trauma and no language skills, that isn't always showing up in the data... It's the social emotional data that I don't see them looking at.

Other principals note that the district is not very familiar with the school specifics: the school program and academic focus, what the school is doing well, and how much the school has improved.

Principal Sarah voices this concern:

I think for an accurate diagnostic, someone would need to know my school on a deeper level, and there would need to be commitment from folks from [the] central office to really spend some time in our school, to understand, and to be able to speak to ... different initiatives that we have, the way that our school culture runs.

Principals reported that test scores remained the dominant information considered. Overall, a majority of principals believe that the diagnosis accuracy could be improved if these were based on mechanisms that provide a close look at individual schools, and not from a system-level evaluation of quantitative metrics. Many principals highlighted the district's incomplete knowledge of schools' challenges and strengths as well as why students fail.

School Inspections: Positive Attitudes among Principals

Principals' attitudes towards Sis can be categorized as positive, mixed, or negative. A summary of findings is presented in Table 5. I assess these attitudes in terms of perceived diagnosis effectiveness, sentiments of appropriateness, and readiness for change.

Perceptions of Diagnosis Effectiveness

A majority of the principals (70%) believe that SIs can effectively evaluate schools. Most of these principals explain that inspection is effective due to the thoroughness of the process. School visits offer "boots on the ground" (Principal David) and provide a "more holistic" diagnosis (Principal Thomas). Principal Monica illustrates this point in more detail:

[The inspectors] actually lived the life of the school ... they were in the building two days; they were in every single classroom. So they had an opportunity to see every single teacher and the fact that they were able to touch every single piece, it meant that they knew everything that was going on in that building ..., I really felt at the end, we had a really good, strong picture of the school.

Table 5. Summary of principals' views

Attitudes	Diagnosis Effectiveness	Appropriateness	Readiness to change
<u>Positive</u>	Accurate knowledge of schools Holistic Thorough Contrasting different sources "Real"	Aligned with school vision & long-term goals Opportunity to assess areas of interest Insights from staff Confirms what schools are doing well Source of legitimacy to implement reforms Enables smooth changes	Strong commitment Changes based on convictions No sense of obligation Perceptions of change efficacy Already have evidence that changes worked
<u>Ambivalent</u>	Limits to know the school in a short visit Limits of any external review	Disagreement with feedback Disagreement on "high leverage" areas Disagreement on specific feedback Already implemented different plans Discussed other changes with district leaders Staff should decide handling of specific issues	Some doubts about self-efficacy Not sure school staff will "buy-in" Staff tired of constant reforms Changes are not central for the district
<u>Negative</u>	Inability to know the "true culture" of schools	Doesn't focus on what is important for the school	Questions towards the whole process validity Negative tone of evaluations

Principals highlighted the value of different mechanisms used during the school visit for attaining an accurate diagnostic. Interestingly, there was no consensus regarding which mechanisms are the most valuable. Principals mentioned varied perspectives about how the varied sources of information led to a thorough knowledge of the school. For example, Principal Mary valued inspections as an “honest insight to what was actually going on inside the classrooms,” referring to classroom observations. Principal Nicholas explained that through focus groups and interviews they “learned a lot about just how ... [the teachers] were feeling and how... [school leaders] could help with their morale.”

Principal Monica thought that by analyzing both school data and school planning documents prior to the visit, the inspectors “walked in the door,” and are able to say: “Yeah. This is exactly what it looks like.” Despite differing in their perception of the most valuable information sources, principals highlight how these sources can reveal different aspects of the schools.

Several principals draw direct comparison between the SIs with the district accountability framework to illustrate their points of view regarding perceived accuracy. Principal Amy notes that the inspectors “come into our space to ask the questions, whereas [with the district accountability framework], you look at a school on paper. You can't read between the lines [about] a school on paper.” Principal David explains that “the outer inspection really provided ... [an] on the ground assessment [in a way that the accountability framework] just isn't capable of providing.” Principal Thomas mentioned that he felt like the inspectors actually “asked more questions than a lot of people of the district asked.” Similarly, several principals draw direct comparison with the work of their supervisors, who also visit the school and get to see the local context in greater detail. Some principals believe that the inspections provide a more thorough and objective assessment. Other principals acknowledge that their supervisors know the schools well and trust the assessments. Yet, most principals agree that these sentiments are greatly reliant on the individual supervisors.

Overall, most principals deem the SI diagnoses as more effective than the data accountability framework. In comparison to the district supervisors, it is unclear which is perceived as more accurate; there is a difference in opinion among interviewed principals.

Sentiments of Appropriateness

To assess sentiments of appropriateness, the interviews inquired whether reforms based on the SI are aligned with the school vision, programming, and long-term goals. Nearly 70% of principals confirm that the feedback aligns with at least one of these aspects.

Many principals think that the inspection feedback aligns with the programming and vision (56% of respondents). They explain that this alignment is due to the evaluation's broad scope and insights gained from school members (administrators, teachers, and students).

Principals feel that the SI encompasses a wide variety of issues that go beyond the district accountability framework. This broadens the possibilities for informing and strengthening school programming. For example, Principal Matthew's programming has been focusing on teachers' growth mindset, although he feels these efforts have been "falling short." The SI addressed this focus and provided a chance to strengthen it; Principal Matthew has "capitalized" on the SI feedback as an opportunity for collegial learning. He explains: the SI has been "fitting with where the school needs to go" with a "very helpful and time efficient" process.

Principal Linda provides another example. She manages a school with a high proportion of immigrant students. Her school programming addresses trauma and socio-emotional wellbeing; this is viewed as a necessary condition for improving student performance. She explains that the wider scope of the SI is appropriate in multiple ways: it aligns with their programming, affirmed that they "honor the diversity," and identifies areas of strength and improvement.

In addition, several principals emphasized that on-site SI evaluations can obtain critical insights from school staff and students. This wealth of information aligns well with the schools'

values and has been used to inform, support, and adjust the schools' programming. Principal Nicholas provides an example about key insights helping to pursue their programming:

I think [the SI] has been very supportive of our overall vision as a school. Our vision is to empower students to be self-agents, to be independent, to have a voice. And teachers don't seem to exercise enough of their [own voice] ... the fact that it came out in the SI that they felt powerless, that they felt like they had no voice in decision-making and decisions that affected them, really shook me ... That's why we made such a huge change very quickly to address their concern.

I also assess the alignment between SI feedback and long-term goals. To do so, all changes in the school vision, planning documents, and explicit mentions to the in the interviews about the long-term goals were considered. About 56% of respondents believe there is alignment between the SI and long-term goals of the school. Three of the interviewed principals based their school's vision on SI feedback. One of these principals, Principal Tyler, illustrates this modification:

I think it [the SI] laid the groundwork for us to start to have [a vision] We really didn't have a school vision or mission, or we didn't have a strategic way of attack of what we're doing. So I think that helped uncover we needed to have it.

Some principals mention that the SI aligns with their long-term goals better than the district accountability framework. For example, Principal Nicholas explains that the SI focuses on teacher voice, rather than performance data, fits with their "values and vision as a school." He continues: "That's why I found it very encouraging, because they did focus more on the people and on the long-term rather than the short-term gains." A similar point is made by Principal Thomas:

[The district accountability] framework freaks me out because, sometimes, when I feel like I'm going to focus on the long-term, then I don't think that it will necessarily show up on [the framework]. And then I think I'm going to get lots of questions about whether I'm a good leader or whether my school is going in the right direction. I think the SI has validated that

we're doing really good things here [at our school] and I would like to have more holistic tools to be able to showcase the work that's happening here.

Other principals appreciated the discretion that they have to make gradual changes; this enables them to maintain long-term goals. Principal Mary illustrates this point saying that the SI did not say they had to “make some radical changes”; on the contrary, it said: “here are some strategies” that you can we implement with fidelity that will make the greatest impact in your classroom, in our school, school-wide.”

Readiness for Change

The analysis of readiness for change emphasizes two elements: *commitment to change* and *change efficacy* (the sentiment that a change will be successfully implemented). About 75% of principals expressed their commitment for reforms suggested by SI. In interviews, principals were directly asked about their commitment to these changes. In their responses to this direct question, most principals conveyed a strong commitment. At various points in interviews, principals expressed their degree of commitment. In most cases, principals’ commitment appeared to be based on their own convictions, and not as an obligation. For example, Principal Tyler said that is “super convinced” of the changes and Principal Linda that is “totally committed” and “took to heart” parts of the feedback.

Following Herscovitch and Meyer (2002) typology, I find that most expressions of commitment were either of the *affective* or *continuance* type. Affective commitment arises when principals want to implement the changes, while continuance implies that reforms must be implemented in order to avoid failure. Besides these two motivations, only a few principals acknowledge some *normative type* of commitment, which is spurred by a sense of obligation. This distribution of responses—with a low incidence of normative the *normative* type—is associated with a higher level of commitment (Herscovitch & Meyer, 2002).

Most principals expressed their commitment to specific changes they had already implemented based on the SIs. About 88% of principals had implemented or planned reforms based on the SI at the time of the interview. In a majority of these cases, these changes were part of key reform areas in the schools. This commitment was evident in the three schools where the SI led to establishing a new mission, vision, or strategic plan. This was illustrated by Principal Tyler who has restructured their strategic plan around their four areas that they that the SI focused on, which he considered that “changed the trajectory of the school.” In other cases, the SI led to inclusion of a more targeted focus or new areas for reform. Principal Sarah illustrates expressions the commitment toward these reforms:

We talked about instructional rigor, and that was one that we really latched onto, and that we really took, and thought about, and turned over in our heads, and tried to figure out for the next year, for us to say, "Our school-wide focus, we're going to focus on rigorous instruction," that helped us out.

Additionally, some of the strongest expressions of commitment coincide with perceptions of the SI as a source of legitimacy to carry out school reform plans. This arises in nearly a third of the interviews. These perceptions refer to increased legitimacy with respect to either the school staff and/or the district. Principal Nicholas mentioned both of these perspectives. He implemented “drastic” changes in teacher’ schedules, reducing time devoted to professional development and adding planning time. This decision was based on SI feedback that indicated teachers had low morale and felt their voices were not heard. Principal Nicholas explains how the SI served as a source of legitimacy to move forward with this reform:

I'm completely committed. [This initiative] was really my idea. I really wanted this to happen long before. I didn't think it was possible this kind of change. I didn't think the District would allow us, but because the sentiment came across so strongly in the SI I felt like we had a lot stronger case to present to the District to say, "Hey, this is what the school wants, the staff

is asking for. It came across in the SI as a huge concern and a challenge and we have all this staff input into this plan." And so, I think all that helped to push the District on a decision that I thought would've been impossible without all of that work from the SI to the staff input So, where we might've felt tempted to just do away with it, now there's just a lot more excitement to continue with it.

The interviews asked about principals' perceived efficacy of reforms due to the SI. Half of principals among those who express commitment to change, also stated about change efficacy always favorably. In most of these cases, confidence in efficacy was based on what they observed in previously implemented changes. This was the case of Principals Linda and David:

Principal Linda: I'm totally committed ... Because I've seen a change in the students and in the teachers and in the number of suspensions. All of that has changed.

Principal David: we were able to really ... use that [(the SI report)] to dig into our formative assessment practices. And then I think we ended up seeing a lot of really strong results out of that.

This subset of principals shows a strong readiness for changes based on the SIs, even when there are specific aspects of the feedback that they do not embrace or are not certain about the efficacy. The next section describes these instances of mixed sentiments towards changes.

Mixed Sentiments and Ambivalence

While most principals show overall positive attitudes towards the SI diagnosis, feedback, and readiness for changes based on it, many also raise some concerns about at least one of these categories (see Table 4).

Concerns about Diagnosis Effectiveness

Specific concerns raised about diagnosis effectiveness, were not emphasized. Reasons behind these concerns varied considerably. Two types of concerns were mentioned in several interviews: 1)

the capacity to understand schools during a short visit, and 2) general limitations of an external review to make an effective diagnosis.

First, a few principals questioned the ability of the SI to understand the school in the short time of the visit. Principal Matthews, while appreciating insights from the SI, explains that after a short visit “calling out” behavioral management in the school “doesn't have a lot of validity.” Similarly, Principal David thinks that while specific aspects of the diagnostic were accurate, others were not for the following reason: “I didn't feel it was the most helpful ... around some “soft skills” or areas around culture and climate ... I didn't think that in the two days they were really able to pick up a lot of necessary context and know the best direction for us to go.”

Second, two principals raise general concerns about the limitations of diagnostic tools that are external to the school. Principal Linda mentioned the lack of emphasis placed on the social emotional aspects of a school, arguing that an improvement would be to assess the distress of the students. In this regard, she claims that this is not captured by the district accountability framework, district leadership, nor the SI. Principal Sarah perceives that there are different “pockets of data,” but nobody can tell the full story. Thus, there are principals who feel that neither the SI or the school district can capture school conditions in a comprehensive way.

Concerns about Appropriateness

Among the principals who find the SI to be appropriate overall for their school, half of them identify specific aspects that are not well-suited. General sentiments of appropriateness of the feedback does not translate in buy-in of every aspect included in the inspection. The most common reason for concern is that principals have different focus or do not think is the best path to take or “high leverage” (Principal Thomas). This stance is also illustrated by Principals Matthew:

A couple years ago, we had [an inspection that] ... focused a lot on behavior ... and the umbrella of culture. To me, the issue wasn't around anything but expectations. The expectations were too low ... I pretty much ignored it. I continued to focus on making sure

that we have standard based instruction, to make sure that we have high expectations. That was my means to improve culture.

In a couple of cases, principals thought that some areas of the SI feedback did not align with their strategic planning that the school had decided or that has already been discussed with the school supervisor. This was illustrated by David:

Principal David: [What I discussed with my supervisor] didn't always match with what the inspection surfaced as the biggest gaps. And at times it got a little tricky to try to balance those two and saying: "this is what I know our school needs, and this is what our team has decided." ... [Because] there's only so much that you can have the capacity to change at a certain time. ... I think that led to some in-depth conversations with my [supervisor].

Similarly, Principal Thomas decided not to focus on specific SI feedback and instead relied on the professional decisions of his staff:

One of the recommendations was around better structures related to lesson planning. ... with a really professional staff who's taught for a long time, we chose not to focus on that because that felt constraining to some of our staff members ... they have their own way to plan.

Uncertainty about Readiness for Change

Among principals who were committed to the changes based on the SI, some expressed uncertainty about reform efficacy. In general, prevalence ambivalence feelings about the changes was very limited. The major source of uncertainty about the success of the changes based on the SI, mostly because of the challenge it poses:

Principal Matthew: [I am] ... about 70 percent [confident that these changes can be successfully implemented]. ... Because we're a hard school. ... To have positions that are half in the classroom, half out of the classroom [(a reform implemented based on the SI feedback)], it sounds to me like a hell job, and I don't know how it'll work, quite frankly. So it's a little bit of an experiment.

Principal Heather: “When I first saw the [SI] report, I wasn't sure how much we could really get done quickly, because there was a lot that [was] needed. And when I was hired it was already June, so teachers had left, and we were really scrambling to pull a team together ... and not lose school culture while we did these changes.

Several principals expressed more concerns about whether the school community would buy in the changes. This stance was illustrated by Principals David and Heather:

Principal David: We had some people who are ready to jump in right away. Some people who took more time, especially teachers who had been at the school for a while, I think it's going to felt more in that line of, we've been here, we've ridden this roller coaster before, how are we going to drop and then go back up that sort of thing.

Principal Heather: It took a good share of one entire school year for people to completely buy into this. And the same was true of students, they weren't necessarily loving some of the changes either, because they felt like some of their freedoms had been taken away.

The lack of centrality of the SI as an accountability and improvement mechanism in the district is another factor that seem to increase ambivalence towards changes. The district accountability framework and supervisors' recommendations are the central mechanisms. This is the case of Principal Amy; she found that the SI had an accurate diagnostic, appropriate feedback, and she was committed to the proposed reforms. However, she noted that the SI “doesn't feel central.” Similarly, Principal Sebastian, who acknowledges the effort to make out the SI and find some specific aspects useful, still considers the SI secondary: “I felt like I should honor that work ... but it was not as important to me as my district leaders' feedback.” Finally, when Principal Linda was asked whether the SI facilitates the implementation of long-term goals, she explained that it gave her some ideas, but “there's district mandates and they don't always align with what the schools can do.”

Despite principals' positive attitudes toward the SI, and their commitment stemming from their own convictions, some principals feel pressure to improve. Several principals that show overall positive attitudes still perceived the process to be judgmental at times (Principal Mathew); some felt "like you were under the gun" (Principal Sarah). Principal David explains how he experiences pressure:

Pressure of leading a red school [(the lowest performing school in the district accountability framework)], it's very much implicit ... you get the SI, you get this big assessment that tells you everything's wrong with your school that does, I think in any principal really provide a source of motivation and pressure to want to improve."

Despite the pressure, explicit or implicit, when most of these principals did not deem SI reforms to be appropriate or a priority, they maintained their previous focus and ignored the feedback. Despite these decisions, these group of principals deemed the feedback overall as appropriate and are prepared to implement changes.

Negative Attitudes

A group of four principals hold attitudes contrary to lasting change towards the whole SI process. The reasons underlying these attitudes are varied (see Table 4). Interestingly, comments about the diagnosis effectiveness were not as prevalent. Only two principals directly question the accuracy and the capacity to know the school. Principal Sebastian argues that he "didn't feel like [the SI] observed all of the important systems" they had in place; however, he does not question the overall diagnosis effectiveness. Principal Brian poses the strongest criticism, directly questioning SI capacity to become familiar with schools, due to the short visit and lack of follow-up:

I didn't feel [they] understood our true culture and our school. The amount of time spent... there's got to be an ongoing process where they're spending more time in our schools, more time understanding true context, just getting follow-up, after the SI is completed, for support

... [The SI resources] would have been invaluable; but somebody coming in for a couple of days ..., and then getting back on the plane, that's not helpful.

In contrast, all principals in this group were critical of the SI feedback not being appropriate for their school. The basis for these attitudes varies. Several principals criticize the focus—or lack of focus—of the evaluation instrument. For example, Principal Ashley characterized the SI as “just a compilation of best practices that you would expect to see in pretty much any school across the planet, [that] doesn’t say anything innovative or riveting.” Similarly, Principal Paul argue that the SI is “only interested in a handful of items that never really took our school-wide focus into account.” Principal Brian emphasizes the lack of usefulness: “I don't think they told us anything that we didn't already know. They didn't give us anything that led directly to anything that we're currently doing instructionally and culturally.”

Three principals expressed their lack of commitment to the SI process. They dismissed the overall validity of the SI process. This is illustrated by Principals Paul and Brian:

Principal Paul: They came into the building for about three days, they used the room to conduct interviews, they went to multiple classrooms to take notes, and they formulated their own report and gave it back to us. So, it didn't seem authentic, I guess.

Principal Brian: I didn't find it a very useful process or tool. You know, I just didn't feel like it was authentic. ... I wouldn't put a lot of weight into the impact it had. It was a report that I probably read one time and that was it and then we moved on.

Accountability pressure and decision-making control appears as underlying themes across most interviews. These perceptions might influence attitudes towards SIs. Interestingly, three out of the four principals who demonstrated strongly negative attitudes only participated in the process once, during the one year that the SI had high-stakes. In contrast, among principles with positive attitudes towards SIs, only one out of 12 participated solely during the year when high-stakes were in effect. In all of these cases, there is a perception of high pressure and a feeling that they were required to

implement recommended reforms. For example, Principal Ashely say that she is not going to “disrupt the school improvement cycle of our entire school based upon one metric.” Principal Paul remarked about the lack of transparency for evaluating quality:

There was a lack of transparency. They didn't want to talk about the questions that they wanted to ask teachers or students. ... They didn't provide a rubric for how things were scored. ... Say what it is that you're going to look for in a SI “Here is the scoring system in which we are going to use this inspection”

Similarly, this group of principals also mentioned that the inspection felt evaluative and highlighted negative aspects of the process. Further, there is a sense of perceived unfairness. Principal Sebastian explains that “the process was informative”, yet he “also felt vulnerable as the principal.” Principal Paul say that the SI “wasn't about guiding feedback for a development of instructional learning. It was more almost a negative, these are the things that you're not doing.” Principal Ashley voices her frustration in a more personal tone: “I just find it belittling that individuals think the external tool is going to make us somehow improve our buildings when it's something that we do every single day of our lives as public educators.” She goes a step further, questioning the purpose of the SI: “I believe the SI is a passive tool to confirm the district's perception of a school so that they can ... support their decisions that they already have of closing schools.”

Overall, while the basis for negative attitudes varies, it leads principals to dismiss the validity of the process. Principals with negative attitudes do not perceive the feedback as providing useful information. They discard most results, in part due to feelings of accountability pressure and a sense of obligation to make changes based on the feedback.

Conclusions

Inspections may offer an alternative to test-based accountability to help schools gain insight into promising directions for improvement. Results indicate that the majority of interviewed principals have positive attitudes towards inspection. The positive attitudes demonstrated by

principals are ones associated with lasting institutional change; this suggests that inspection might enable sustained school reforms. In addition to inspection, all schools in the study also are subject to high-stakes accountability that emphasizes standardized testing. Yet, a majority of principals questioned the ability of testing alone to effectively evaluate their school. Test scores alone do not provide explanations for low performance or provide limited directions for specific reforms.

A major contribution of this study is that it demonstrates the connection between positive attitudes and known dispositions that lead to lasting change. It finds that most principals perceive the diagnosis as effective, feel the suggested changes to be appropriate, and are ready to take action. These results are consistent with previous studies that find principals have positive attitudes regarding inspections (Behnke & Steins, 2017; Bitan et al., 2014; C. Gray & Gardner, 1999). This study goes beyond past efforts, focusing on attitudes associated with sustained change.

Principals appear to form positive attitudes based on aspects of inspection that are absent from test-based accountability. Perceived effectiveness of the diagnosis is attributable to the thoroughness of the inspection process. Inspections result in an accurate picture of the school and identify key challenges, strengths, and improvement areas. Perceived appropriateness of the feedback is associated with the SI considering a comprehensive set of reform areas that inform school planning. Many principals view SI findings as an opportunity to support and refine their plans for the school to achieve long-term goals. Finally, readiness for changes recommended in the inspection feedback is evident; principals expressed commitment to these improvement areas. In addition, they show great confidence that reforms will be effectively implemented.

However, most principals also expressed some ambivalent feeling towards aspects of inspection. Some argued that brief, external evaluations can be limited in their ability to accurately assess school conditions. Others felt that the specific feedback did not align well with their strategic planning, or that it undermined the decision-making making power of school staff. In addition, several principals have some doubts on the efficacy of changes, recognizing that change is hard, and

reforms can fail. Overall, most principals have strong positive attitudes towards inspection, yet some note the limitations of certain aspects.

Principal attitudes appear to be associated with accountability pressure and perceived control over decision-making. When principals perceive greater accountability pressure, they are more critical of the process and less likely to implement recommended reforms. This was evident for a small number of principals who only received one inspection visit during a year when stakes were higher. While specific criticisms varied, these principals questioned the validity of the overall process and motives for the inspections. Yet, most schools had at least one inspection in years with lower accountability pressure and only faced an implicit pressure to improve. Most principals do not feel obliged to implement changes based on the results. Principals are selective and target feedback that is useful to support their long-term plans.

This paper sheds lights on how principals perceive inspection. I contrast these perceptions with those of district's main mechanism for accountability: test-based accountability and school supervisors. Test-based accountability is viewed as a transparent way to measure school performance, yet it is limited in its ability to identify specific strengths and weaknesses. Then, principals' supervisors are familiar with the local context and school practices; this offers the possibility of open dialogues on improvement strategies. Yet, principals view that individual assessments by supervisors are not as transparent as inspections. In addition, this in-depth knowledge relies on specific individuals, and risks being lost. In contrast, inspections appear to offer improved transparency and insights for reforms. On-site evaluations evaluate school operations following a protocol and produce formal reports; the process is clearly defined and principals find the assessments useful.

Lastly, principals highlighted limitations of the district inspection policy to inform school reforms. First, the lack of centrality of inspection in contrast to the district accountability framework. Principals have strong incentives to respond to district ratings, which are heavily based on

standardized test results. They do not exhibit the same urgency to implement changes based on inspection feedback. Second, many principals feel that the lack of follow-up procedures after inspection limits the ability of the district to support reforms they implement in response to inspection feedback.

Overall, this study presents evidence of the potential for school inspection to enable sustained reforms in systems dominated by high-stakes accountability. It shows that brief inspection visits are perceived by principals as effective and provides motivation to implement reforms based on feedback. It shows the value of evaluating schools holistically and considering school stakeholder' perspectives, to provide insight for improvement that is well-aligned with school goals.

APPENDICES

Appendix A. Interview Protocol

IRB application ID#: STUDY00001267

SECTION I – DISTRICT DIAGNOSIS AND KNOWLEDGE OF THE LOCAL CONTEXT

1. First, tell me a bit about your background in education. How did you come to be a principal?
2. How much do you think the District Office knows about your school and its context?
 - a. If the district decided to do an integral diagnosis of your schools with the information they have now, how accurate do you think it would be?
 - b. What sources of information would the district use for this diagnosis (considering information the district already has)?
 - c. What information would the district be missing to make an accurate diagnosis of your school?
 - d. Would you consider this the district diagnosis to be fair?
3. What feedback regarding school performance or how well your school works have you received from the district?
 - a. Which of those were most useful to you? Why?
 - b. To what degree can you (or you and a school-based committee) determine or interpret what changes should be implemented in your schools based on this feedback?

SECTION II – SCHOOL IMPROVEMENT AND DECISION MAKING

4. Can you tell me about an important area of your school you are currently or have recently been working to improve?
 - a. Why did you select FILL IN as the area to focus on?
 - b. Thinking back, how did you select FILL IN as a focus for improvement?
 - i. Were there any resources, including people, that helped inform your decision?

- c. What information did you examine prior to selecting FILL IN as an area to focus on?
- d. IF NOT ALREADY SAID: What have you and your faculty been doing to address this area?
 - i. What resources, including people, helped you decide this course of action?
- 5. Looking back, is there any information or resources you wished you had to help you plan strategically for school improvement?
- 6. What is the most important the district feedback or support mechanism that led you to change how you see the main problems in your school?

FOLLOW-UP:

- Performance framework, school supervisor, and SI.

SECTION III – SCHOOL INSPECTION (SI): PERCEPTIONS & RESPONSE

- 7. Can tell me about the time you received an SI visit in your school?
 - a. How it was the process and how did you personally experience it?
 - b. In what ways was this process useful for you?

FOLLOW-UP:

- To what degree has the SI led you to change how you “see” the main problems that hinder improvement in your school?
- Has the SI helped you become aware of any significant changes were needed at your school in order to improve education quality?
- To what degree did SI feedback confirm what you already suspected?
 - If the SI feedback confirmed what you already now, was it helpful to have this confirmation? Why?
- In what ways it wasn’t helpful enough.

8. Do you think that the changes that should be implemented based on the SI feedback are aligned with your school history, vision, and current programming?
 - a. How are the changes that should be implemented based on the SI feedback more or less aligned with your values and current programming in comparison to the changes based on other the district assessment and support mechanisms?
[if not mentioned, ask about the Performance Framework and the School Supervisor]
 - b. How convinced are you that that these changes can be successfully implemented?
 - c. How motivated and committed are you to these changes? Why?
Follow-up: To what extent are you motivated because you feel these changes are valuable/ appropriate vs. because you are required to carry these out (& could face consequences)?
 - d. How invested was your staff in following these changes?
9. Can you tell me how the most relevant changes on focus or strategy you implemented based on the SI feedback? Can you give me an example of how you used the information in the SI to create change?
 - a. Why have you decided to focus on FILL IN?
 - b. How was the SI useful in implementing this change?

FOLLOW UP:

Was the SI helpful to define specific improvement strategies? How?

What was the role of the school supervisor in this process?
10. What significant change could have been implemented based on the SI feedback, but you decided not to?
 - a. What circumstances led to these changes not being implemented?
11. Due to accountability pressure to improve, have you ever considered changing strategies that you think would deliver improvements in the long-term, but not in the short-term?

- a. Have the SI changed this idea in any way?
12. I'm sure you spend a lot of time reviewing data and other indicators of your school's quality. How do you sort through all of this information? Can you walk me through how you approach the different sources of data and how you weigh different sources?
- a. Compared to FILL IN, how useful is the SI feedback for the school improvement planning process?
 - b. Compared to FILL IN, how useful is the school supervisor for the school improvement planning process?
 - c. Compared to FILL IN, how useful are School Performance Framework for the school improvement planning process?
 - d. In what ways is the SI information different from other sources of school quality indicators?
13. Overall, how valuable do you think the SIs are? How could they be improved?

SECTION IV – PRINCIPAL BACKGROUND

14. How many years total have you been a teacher?
15. PRIOR to this school year, how many years did you serve as the principal of THIS OR ANY OTHER school? (Count part of a year as 1 year)
16. PRIOR to this school year, how many years did you serve as the principal of THIS school? (Count part of a year as 1 year)
17. What is the highest degree you have earned? IF NOT STATED, Is that degree specific to education leadership?

Appendix B. Coding Scheme

1. Diagnosis accuracy / knowledge of school

1.1. Accurate / Good knowledge

1.1.1. Inspection

1.1.2 District

1.2 Inaccurate / Incomplete knowledge

1.2.1. Inspection

1.2.2 District

2. Diagnosis usefulness

2.1. Useful / New insights

2.1.1. Better prioritize

2.1.2. Reaffirm existing goals / Confirm diagnosis

2.1.3. Gain legitimacy

2.1.4. Somewhat useful

2.2. Not useful / Not relevant

3. Appropriateness

3.1 Most Appropriate

3.1.1. Aligned with school history, vision, values

3.1.2. Focus on the long term

3.1.3. Consistent with the District

3.2. Not a priority

3.3. Not appropriate

4. Commitment / Trust in efficacy

4.1. Commitment / Trust in efficacy

4.2. Ambivalent attitudes

4.3. Lack of commitment / Distrust efficacy

5. Changes – What motivates them?

5.1. Principal initiatives / Staff initiatives

5.2. Inspections

5.3. Test results, performance framework, evaluations

5.4. School supervisors

5.5. District (excluding 5.3. & 5.4)

5.6. Other sources

5.7. Did not implement changes based on inspection (explicit)

6. Principals views

6.1. Critical views

6.1.1. Inspections

6.1.2. District

6.1.3. School supervisors

6.2. Positive views

6.2.1. District

6.2.2. School supervisors

REFERENCES

REFERENCES

- Ahuvia, A. (2001). Traditional, Interpretive, and Reception Based Content Analyses: Improving the Ability of Content Analysis to Address Issues of Pragmatic and Theoretical Concern. *Social Indicators Research*, 54, 139–172. <https://doi.org/10.1108/781350505>
- Allen, R., & Burgess, S. (2012). *How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England* (No. 12; 87).
- Altrichter, H., & Kemethofer, D. (2015). Does accountability pressure through school inspections promote school improvement? *School Effectiveness and School Improvement*, 26(1), 32–56. <https://doi.org/10.1080/09243453.2014.927369>
- Apple, M. (2005). Education, markets, and an audit culture. *Critical Quarterly*, 47(1–2), 11–29. <https://doi.org/doi:10.1111/j.0011-1562.2005.00611>
- Armenakis, A., Bernerth, J., Pitts, J., & Walker, H. (2007). Organizational Change Recipients' Beliefs Scale. *The Journal of Applied Behavioral Science*, 43(4), 481–505. <https://doi.org/10.1177/0021886307303654>
- Armenakis, A., & Harris, S. (2009). Reflections: our Journey in Organizational Change Research and Practice. *Journal of Change Management*, 9(2), 127–142. <https://doi.org/10.1080/14697010902879079>
- Armenakis, A., Harris, S., Cole, M., Fillmer, L., & Self, D. (2007). A Top Management Team's Reactions to Organizational Transformation: The Diagnostic Benefits of Five Key Change Sentiments. *Journal of Change Management*, 7(3–4), 273–290. <https://doi.org/10.1080/14697010701771014>
- Armstrong, J. (1982). The value of formal planning for strategic decisions: review of empirical research. *Strategic Management Journal*, 3, 197–211.
- Ball, S., & Bowe, R. (1992). Subject departments and the 'implementation' of National Curriculum policy: an overview of the issues. *Journal of Curriculum Studies*, 24(2), 97–115. <https://doi.org/10.1080/0022027920240201>
- Barber, M. (2005). The virtue of accountability: System redesign, inspection, and incentives in the era of informed professionalism. *Journal of Education*, 185(1), 7–38. <https://doi.org/10.1177/002205740518500102>
- Baxter, J. A. (2013). Professional inspector or inspecting professional? Teachers as inspectors in a new regulatory regime for education in England. *Cambridge Journal of Education*, 43(4), 467–485. <https://doi.org/10.1080/0305764X.2013.819069>
- Behnke, K., & Steins, G. (2017). Principals' reactions to feedback received by school inspection: A longitudinal study. *Journal of Educational Change*, 18(1), 77–106.

<https://doi.org/10.1007/s10833-016-9275-7>

- Bengston, D., & Xu, Z. (1995). *Changing National Forest Values: a content analysis - Research Paper NC-323*. http://www.nrs.fs.fed.us/pubs/rp/rp_nc323.pdf
- Berry, F. S., & Wechsler, B. (1995). State agencies' experience with strategic planning: findings from a national survey. *Public Administration Review*, 55(2), 159.
<https://doi.org/10.2307/977181>
- Bitan, K., Haep, A., & Steins, G. (2014). School inspections still in dispute – an exploratory study of school principals' perceptions of school inspections. *International Journal of Leadership in Education*, 18(4), 1–22. <https://doi.org/10.1080/13603124.2014.958199>
- Bloem, S. (2015). The OECD Directorate for Education as an independent knowledge producer through Pisa. In H. G. Kotthoff & E. Klerides (Eds.), *Governing Educational Spaces* (pp. 169–185). SensePublishers. https://doi.org/10.1007/978-94-6300-265-3_10
- Brier, A., & Hopp, B. (2011). Computer assisted text analysis in the social sciences. *Quality & Quantity*, 45(1), 103–128. <https://doi.org/10.1007/s11135-010-9350-8>
- Chabbott, C., & Elliott, E. J. (2003). Understanding others, educating ourselves: Getting more from international comparative studies in education. In *Social Sciences*.
<https://doi.org/10.17226/10622>
- Chun, Y. H., & Rainey, H. G. (2005). Goal ambiguity and organizational performance in U.S. federal agencies. *Journal of Public Administration Research and Theory*, 15(4), 529–557.
<https://doi.org/10.1093/jopart/mui030>
- Clarke, J., & Ozga, J. (2011). Governing by inspection? Comparing school inspection in Scotland and England. *Social Policy Association Conference*, 25.
- Coburn, C. (2001). Beyond decoupling: Rethinking the relationship between the institutional environment and the classroom. *Sociology of Education*, 77, 211–244.
<https://doi.org/10.1177/003804070407700302>
- Coburn, C. (2005). Shaping teacher sensemaking: School leaders and the enactment of reading policy. *Educational Policy*, 19(3), 476–509. <https://doi.org/10.1177/0895904805276143>
- Cole, M. S., Harris, S., & Bernerth, J. B. (2006). Exploring the implications of vision, appropriateness, and execution of organizational change. *Leadership & Organization Development Journal*, 27(5), 352–367. <https://doi.org/10.1108/01437730610677963>
- Concurso de Supervisores Rio Negro, (2013).
- Resolución del Consejo Provincial de Educación de Río Negro N 1053, Pub. L. No. 1053 (1994).
- Conway, M. (2006). The Subjective Precision of Computers: A Methodological Comparison with Human Coding in Content Analysis. *Journalism & Mass Communication Quarterly*,

- 83(1), 186–200. <https://doi.org/10.1177/107769900608300112>
- Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Cuckle, P., Hodgson, J., & Broadhead, P. (1998). Investigating the relationship between OFSTED Inspections and school development planning. *School Leadership & Management, 18*(2), 271–283. <https://doi.org/10.1080/13632439869691>
- Darling-Hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Podolsky, A., & Stosich, E. L. (2016). *Pathways to new accountability through the Every Student Succeeds Act*. <http://learningpolicyinstitute.org/our-work/publications-resources/%0Apathways-new-accountability-every-student-succeeds-act>.
- De Vries, H., Elliott, M. N., Kanouse, D. E., & Teleki, S. S. (2008). Using Pooled Kappa to Summarize Interrater Agreement across Many Items. *Field Methods, 20*(3), 272–282. <https://doi.org/10.1177/1525822X08317166>
- de Wolf, I., & Janssens, F. (2007). Effects and side effects of inspections and accountability in education: An overview of empirical studies. *Oxford Review of Education, 33*(3), 379–396. <https://doi.org/10.1080/03054980701366207>
- Dedering, K., & Müller, S. (2011). School improvement through inspections? First empirical insights from Germany. *Journal of Educational Change, 12*(3), 301–322. <https://doi.org/10.1007/s10833-010-9151-9>
- Dedering, K., & Sowada, M. G. (2017). Reaching a conclusion—procedures and processes of judgement formation in school inspection teams. *Educational Assessment, Evaluation and Accountability, 29*(1), 5–22. <https://doi.org/10.1007/s11092-016-9246-9>
- Deng, Q., Hine, M., Ji, S., & Sur, S. (2019). Inside the black box of dictionary building for text analytics: a design science approach. *Journal of International Technology and Information Management, 27*(3), 119–159.
- Doud, J. (1995). Planning for school improvement: A curriculum model for school based evaluation. *Peabody Journal of Education, 70*, 175–187.
- Edgerton, A. K. (2019). The essence of ESSA: More control at the district level? *Phi Delta Kappan, 101*(2), 14–17. <https://doi.org/10.1177/0031721719879148>
- Education Inspectorate - Ministry of Education, C. and S. (2010). *Risk-based Inspection as of 2009 - Primary and Secondary Education*.
- Education Inspectorate - Ministry of Education, C. and S. (2017a). *Inspection framework primary education*.
- Education Inspectorate - Ministry of Education, C. and S. (2017b). *Inspection framework secondary education*.

- Ehren, M. (2016a). *Methods and modalities of effective school inspections* (M. Ehren (ed.)). Springer International Publishing. <https://doi.org/10.1007/978-3-319-31003-9>
- Ehren, M. (2016b). *Methods and Modalities of Effective School Inspections*. In M. C.M. Ehren (Ed.), *Methods and Modalities of Effective School Inspections*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-31003-9>
- Ehren, M., Altrichter, H., McNamara, G., & O'Hara, J. (2013). Impact of school inspections on improvement of schools—describing assumptions on causal mechanisms in six European countries. *Educ Asse Eval Acc*, 25, 3–43. <https://doi.org/10.1007/s11092-012-9156-4>
- Ehren, M., Gustafsson, J.-E., Altrichter, H., Skedsmo, G., Kemethofer, D., & Huber, S. (2015). Comparing effects and side effects of different school inspection systems across Europe. *Comparative Education*, 51(3), 375–400. <https://doi.org/10.1080/03050068.2015.1045769>
- Ehren, M., Perryman, J., & Shackleton, N. (2015a). School Effectiveness and School Improvement. *School Effectiveness and School Improvement - An International Journal of Research, Policy and Practice*, 26(2), 296–327.
- Ehren, M., Perryman, J., & Shackleton, N. (2015b). Setting expectations for good education: how Dutch school inspections drive improvement. *School Effectiveness and School Improvement*, 26(2), 296–327. <https://doi.org/10.1080/09243453.2014.936472>
- Ehren, M., & Shackleton, N. (2016). Risk-based school inspections: impact of targeted inspection approaches on Dutch secondary schools. *Educational Assessment, Evaluation and Accountability*, 28(4), 299–321. <https://doi.org/10.1007/s11092-016-9242-0>
- Ehren, M., & Visscher, A. (2006). TOWARDS A THEORY ON THE IMPACT OF SCHOOL INSPECTIONS. *British Journal of Educational Studies*, 54(1), 51–72. <https://doi.org/10.1111/j.1467-8527.2006.00333.x>
- Ehren, M., & Visscher, A. (2008). THE RELATIONSHIPS BETWEEN SCHOOL INSPECTIONS, SCHOOL CHARACTERISTICS AND SCHOOL IMPROVEMENT. *British Journal of Educational Studies*, 56(2), 205–227. <https://doi.org/10.1111/j.1467-8527.2008.00400.x>
- Fernandez, K. E. (2011). Evaluating school improvement plans and their affect on academic performance. *Educational Policy*, 25(2), 338–367. <https://doi.org/10.1177/0895904809351693>
- Figlio, D., & Loeb, S. (2011). School accountability. In *Handbook of the Economics of Education* (pp. 383–421).
- Fitchett, P., & Heafner, T. (2010). A national perspective on the effects of high-stakes testing and standardization on elementary social studies marginalization. *Theory & Research in Social Education*, 38(1), 114–130. <https://doi.org/10.1080/00933104.2010.10473418>
- Gagnon, D. J., & Schneider, J. (2019). Holistic school quality measurement and the future of

- accountability: Pilot-test results. *Educational Policy*, 33(5), 734–760.
<https://doi.org/10.1177/0895904817736631>
- Gilroy, P., & Wilcox, B. (1997). OFSTED, criteria and the nature of social understanding: A Wittgensteinian critique of the practice of educational judgement. *British Journal of Educational Studies*, 45(1), 22–38. <https://doi.org/10.1111/1467-8527.00034>
- Gioia, D., Thomas, J., Clark, S., & Chittipeddi, K. (1994). Symbolism and strategic change in academia: The dynamics of sensemaking and influence. *Organization Science*, 5(3), 363–383. <https://doi.org/10.1287/orsc.5.3.363>
- Glazerman, S. (2016). *The false dichotomy of school inspection*. Mathematica Policy Research - Blog Post. <https://www.mathematica-mpr.com/commentary/the-false-dichotomy-of-school-inspections>
- Gray, C., & Gardner, J. (1999). The impact of school inspections. *Oxford Review of Education*, 25(4), 455–468. <https://doi.org/10.1080/030549899103928>
- Gray, J., & Wilcox, B. (1995). In the aftermath of inspection: the nature and fate of inspection report recommendations. *Research Papers in Education*, 10(1), 1–18.
<https://doi.org/10.1080/0267152950100102>
- Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11(13), 255–274.
- Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267–297.
<https://doi.org/10.1093/pan/mps028>
- Grimolizzi-Jensen, C. J. (2018). Organizational change: Effect of motivational interviewing on readiness to change. *Journal of Change Management*, 18(1), 54–69.
<https://doi.org/10.1080/14697017.2017.1349162>
- Gustafsson, J.-E., Ehren, M., Conyngham, G., McNamara, G., Altrichter, H., & O'Hara, J. (2015). From inspection to quality: Ways in which school inspection influences change in schools. *Studies in Educational Evaluation*, 47, 47–57.
<https://doi.org/10.1016/j.stueduc.2015.07.002>
- Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. *Educational Administration, Policy, and Reform: Research and Measurement*, 3, 153–188.
- Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2), 297–327.
<https://doi.org/10.1002/pam.20091>
- Herscovitch, L., & Meyer, J. P. (2002). Commitment to organizational change: Extension of a

- three-component model. *Journal of Applied Psychology*, 87(3), 474–487.
<https://doi.org/10.1037/0021-9010.87.3.474>
- Hill, H. (2001). Policy is not enough: language and the interpretation of State standards. *American Educational Research Journal*, 38(2), 289–318.
<https://doi.org/10.3102/00028312038002289>
- Hines, R. T. (2017). *An Exploration of the Effects of School Improvement Planning and Feedback Systems: School Performance in North Carolina*.
- Holt, D., Armenakis, A., Feild, H., & Harris, S. (2007). Readiness for Organizational Change. *The Journal of Applied Behavioral Science*, 43(2), 232–255.
<https://doi.org/10.1177/0021886306295295>
- Husfeldt, V. (2011). Wirkungen und wirksamkeit der externen schulevaluation; uberblick zum stand der forschung [The impact of school inspection - Does it really work? State of research]. *Zeitschrift Für Erziehungswissenschaft*, 14(2), 259–282.
<https://doi.org/10.1007/s11618-011-0204-5>
- Hussain, I. (2015). Subjective Performance Evaluation in the Public Sector Evidence from School Inspections. *The Journal of Human Resources*, 50(1), 189–221.
- Jacob, B. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5–6), 761–796.
<https://doi.org/10.1016/j.jpubeco.2004.08.004>
- Jones, K., & Tymms, P. (2014). Ofsted’s role in promoting school improvement: the mechanisms of the school inspection system in England. *Oxford Review of Education*, 40(3), 315–330.
- Jones, K., Tymms, P., Kemethofer, D., O’Hara, J., McNamara, G., Huber, S., Myrberg, E., Skedsmo, G., & Greger, D. (2017). The unintended consequences of school inspection: the prevalence of inspection side-effects in Austria, the Czech Republic, England, Ireland, the Netherlands, Sweden, and Switzerland. *Oxford Review of Education*, 43(6), 805–822.
<https://doi.org/10.1080/03054985.2017.1352499>
- Kaplan, S., & Orlikowski, W. J. (2013). Temporal Work in Strategy Making. *Organization Science*, 24(4), 965–995. <https://doi.org/10.1287/orsc.1120.0792>
- Klein, A. (2016). *School inspections offer a diagnostic look at quality*. Education Week.
<https://www.edweek.org/ew/articles/2016/09/28/school-inspections-offer-a-diagnostic-look-at.html>
- Klerks, M. (2012). *The effect of school inspections: a systematic review*. <http://janbri.nl/wp-content/uploads/2014/12/ORD-paper-2012-Review-Effect-School-Inspections-MKLERKS.pdf>
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A

- historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>
- Koretz, D. (2008). *Measuring up*. Harvard University Press.
- Krippendorff, K. (2013). *Content analysis: An introduction to Its methodology* (3rd ed.). SAGE Publications.
- Ladd, H. F. (2016). *Now is the time to experiment with inspections for school accountability*. Brookings. <https://www.brookings.edu/blog/brown-center-chalkboard/2016/05/26/now-is-the-time-to-experiment-with-inspections-for-school-accountability/>
- Ladd, H. F. (2017). NCLB: RESPONSE TO JACOB. *Journal of Policy Analysis and Management*, 36(2), 477–480. <https://doi.org/10.1002/pam.21979>
- Ladd, H. F., & Figlio, D. (2008). School accountability and student achievement. In *Handbook of research in education finance and policy* (pp. 166–182).
- Lee, J., & Fitz, J. (1997). HMI and OFSTED: evolution or revolution in school inspection. *British Journal of Educational Studies*, 45(1), 39–52. <https://doi.org/10.1111/1467-8527.00035>
- Lewin, A. Y., & Minton, J. W. (1986). Determining Organizational Effectiveness: Another Look, and an Agenda for Research. *Management Science*, 32(5), 514–538. <https://doi.org/10.1287/mnsc.32.5.514>
- Lindgren, J. (2015). The front and back stages of swedish school inspection: opening the black box of judgment. *Scandinavian Journal of Educational Research*, 59(1), 58–76. <https://doi.org/10.1080/00313831.2013.838803>
- Luginbuhl, R., Webbink, D., & de Wolf, I. (2009). Do Inspections Improve Primary School Performance? *Analysis*, 31(3), 221–237. <https://doi.org/10.3102/0162373709338315>
- Maitlis, S. (2005). The Social Processes of Organizational Sensemaking. *The Academy of Management Journal*, 48(1), 21–49. <https://doi.org/10.2307/20159639>
- Maitlis, S., & Christianson, M. (2014). Sensemaking in organizations: Taking stocks and moving forward. *The Academy of Management Annals*, 8(1), 57–125. <https://doi.org/10.1080/19416520.2014.873177>
- March, J. G., & Olsen, J. P. (2011). The Logic of Appropriateness. In R. E. Goodin (Ed.), *The Oxford Handbook of Political Science* (pp. 1–22). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199604456.013.0024>
- Mathis, W., & Trujillo, T. (2016). *Lessons from NCLB for the Every Student Succeeds Act*. <http://nepc.colorado.edu/%0Apublication/lessons-from-NCLB>
- Matthews, P., & Sammons, P. (2004). *Improvement through inspection. An evaluation of the*

impact of Ofsted's work. Ofsted.

- Matthews, Peter, Holmes, J. R., Vickers, P., & Corporaal, B. (1998). Aspects of the reliability and validity of school inspection judgements of teaching quality. *Educational Research and Evaluation, 4*(2), 167–188. <https://doi.org/10.1076/edre.4.2.167.6959>
- McDonnell, L. (2008). The politics of educational accountability: Can the clock be turned back? In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability*. Routledge.
- McDonnell, L. (2013). Educational accountability and policy feedback. *Educational Policy, 27*(2), 170–189. <https://doi.org/10.1177/0895904812465119>
- Meyers, C. V., & VanGronigen, B. A. (2019). A lack of authentic school improvement plan development. *Journal of Educational Administration, 57*(3), 261–278. <https://doi.org/10.1108/JEA-09-2018-0154>
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3th ed.). SAGE Publications.
- Millett, A., & Johnson, D. C. (1998). Expertise or “baggage”? What helps inspectors to inspect primary mathematics? *British Educational Research Journal, 24*(5), 503–518. <https://doi.org/10.1080/0141192980240502>
- Mintrop, H., MacLellan, A. M., & Quintero, M. F. (2001). School improvement plans in schools on probation: A comparative content analysis across three accountability systems. *Educational Administration Quarterly, 37*(2), 197–218. <https://doi.org/10.1177/00131610121969299>
- Morse, J. (2010). Procedures and practice of mixed method design - Maintaining control, rigor, and complexity. In A. M. Tashakkori & C. B. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 339–352). SAGE Publications.
- Neuendorf, K. A. (2017). *The content analysis guidebook*. SAGE Publications, Inc. <https://doi.org/10.4135/9781071802878>
- Nusche, D., Braun, H., Halász, G., & Santiago, P. (2014). *OECD Reviews of Evaluation and Assessment in Education: Netherlands 2014*. OECD. <https://doi.org/10.1787/9789264211940-en>
- OECD, O. for E. C. and D. (2015). *Education at a glance 2015 - OECD Indicators*. <https://doi.org/10.1787/19991487>
- Ouston, J., Fidler, B., & Earley, P. (1997). What do schools do after OFSTED school inspections-or before? *School Leadership & Management, 17*(1), 95–104. <https://doi.org/10.1080/13632439770195>
- Penninckx, M., & Vanhoof, J. (2015). Insights gained by schools and emotional consequences of

- school inspections. A review of evidence. *School Leadership & Management*, 35(5), 477–501. <https://doi.org/10.1080/13632434.2015.1107036>
- Penninckx, M., Vanhoof, J., De Maeyer, S., & Van Petegem, P. (2014). Exploring and explaining the effects of being inspected. *Educational Studies*, 40(4), 456–472. <https://doi.org/10.1080/03055698.2014.930343>
- Penninckx, M., Vanhoof, J., De Maeyer, S., & Van Petegem, P. (2015). Effects and side effects of Flemish school inspection. *Educational Management Administration & Leadership*. <https://doi.org/10.1177/1741143215570305>
- Perryman, J. (2007). Inspection and emotion. *Cambridge Journal of Education*, 37(2), 173–190. <https://doi.org/10.1080/03057640701372418>
- Perryman, J. (2009). Inspection and the fabrication of professional and performative processes. *Journal of Education Policy*, 24(5), 611–631.
- Phillips, D., & Schweisfurth, M. (2014). *Comparative and international education - An introduction to theory, methods , and practice* (2nd ed.). Group, Continuum International Publishing.
- Piderit, S. K. (2000). Rethinking resistance and recognizing ambivalence: A multidimensional view of attitudes toward an organizational change. *The Academy of Management Review*, 25(4), 783. <https://doi.org/10.2307/259206>
- Pond, S., Armenakis, A., & Green, S. (1984). The Importance of Employee Expectations in Organizational Diagnosis. *The Journal of Applied Behavioral Science*, 20(2), 167–180. <https://doi.org/10.1177/002188638402000207>
- Porac, J. F., Thomas, H., & Baden-Fuller, C. (1989). COMPETITIVE GROUPS AS COGNITIVE COMMUNITIES: THE CASE OF SCOTTISH KNITWEAR MANUFACTURERS. *Journal of Management Studies*, 26(4), 397–416. <https://doi.org/10.1111/j.1467-6486.1989.tb00736.x>
- Portz, J., & Beauchamp, N. (2020). Educational Accountability and State ESSA Plans. *Educational Policy*, 089590482091736. <https://doi.org/10.1177/0895904820917364>
- Ravitch, D. (2016). *The death and life of the great American school system: How testing and choice are undermining education*. Basic Books.
- Redding, C., & Searby, L. (2020). The Map Is Not the Territory: Considering the Role of School Improvement Plans in Turnaround Schools. *Journal of Cases in Educational Leadership*, 23(3), 63–75. <https://doi.org/10.1177/1555458920938854>
- Riffe, D., Lacy, S., & Fico, F. (2014). *Analyzing media messages: Using quantitative content analysis in research*. Routledge.
- Rigby, J. G. (2015). Principals’ sensemaking and enactment of teacher evaluation. *Journal of*

- Educational Administration*, 53(3), 374–392. <https://doi.org/10.1108/JEA-04-2014-0051>
- Rosenthal, L. (2004). Do school inspections improve school quality? Ofsted inspections and school examination results in the UK. *Economics of Education Review*, 23, 143–151.
- Rothstein, R., Jacobsen, R., & Wilder, T. (2008). *Grading education: Getting accountability right*. Economic Policy Institute and Teacher College Press.
- Rouleau, L. (2005). Micro-practices of strategic sensemaking and sensegiving: How middle managers interpret and sell change every day. *Journal of Management Studies*, 42(7), 1413–1441.
- Rutz, S., Mathew, D., Robben, P., & Bont, A. (2017). Enhancing responsiveness and consistency: Comparing the collective use of discretion and discretionary room at inspectorates in England and the Netherlands. *Regulation & Governance*, 11(1), 81–94. <https://doi.org/10.1111/regg.12101>
- Ryan, K., Gandha, T., & Ahn, J. (2013). School self-evaluation and inspection for improving U.S. schools? In *National Education Policy Center*. <http://nepc.colorado.edu/publication/school-self-evaluation>
- Sandberg, J., & Tsoukas, H. (2015). Making sense of the sensemaking perspective: Its constituents, limitations, and opportunities for further development. *Journal of Organizational Behavior*, 36(S1), S6–S32. <https://doi.org/10.1002/job.1937>
- Scheerens, J., Ehren, M., Slegers, P., & Leeuw, R. de. (2012). *OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes*.
- Shaw, I., Newton, D. P., Aitkin, M., & Darnell, R. (2003). Do OFSTED Inspections of Secondary Schools Make a Difference to GCSE Results? *British Educational Research Journal*, 29(1), 63–75.
- Spillane, J. P. (1999). External reform initiatives and teachers' efforts to reconstruct their practice: The mediating role of teachers' zones of enactment. *Journal of Curriculum Studies*, 31(2), 1–33. <https://doi.org/10.1080/002202799183205>
- Spillane, J. P., Parise, L. M., & Sherer, J. Z. (2011). Organizational Routines as Coupling Mechanisms. *American Educational Research Journal*, 48(3), 586–619. <https://doi.org/10.3102/0002831210385102>
- Spillane, J. P., Reiser, 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 education policy implementation* (pp. 47–64). State University of New York Press, Albany.
- Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of Educational Research*, 72(3), 387–341. <https://doi.org/10.3102/00346543072003387>

- Stiglitz, J. (2000). *Economics of the public sector* (3rd ed.). Norton.
- Strunk, K. O., Marsh, J. A., Bush-Mecenas, S., & Duque, M. R. (2016). The Best Laid Plans. *Educational Administration Quarterly*, *52*(2), 259–309. <https://doi.org/10.1177/0013161X15616864>
- Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating qualitative and quantitative approaches in the social and behavioral sciences*. SAGE.
- Teddlie, C., & Yu, F. (2007). Mixed methods sampling : A typology with examples. *Journal of Mixed Methods Research*, *1*(1), 77–100. <https://doi.org/10.1177/1558689806292430>
- UNESCO. (2017). *Global Education Monitoring Report - Accountability in education: Meeting our commitments*.
- van Bruggen, J. C. (2010). *Inspectorates of Education in Europe; some comparative remarks about their tasks and work*.
- van der Sluis, M. E., Reezigt, G. J., & Borghans, L. (2017). Implementing New Public Management in Educational Policy. *Educational Policy*, *31*(3), 303–329.
- Vavrus, F. K., & Bartlett, L. (2016). *Rethinking case study research: A comparative approach* (1st ed.). Routledge.
- Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2010). Using school performance feedback: perceptions of primary school principals. *School Effectiveness and School Improvement*, *21*(2), 167–188. <https://doi.org/10.1080/09243450903396005>
- Visscher, A. J., & Coe, R. (2003). School performance feedback Systems: conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement*, *14*(3), 321–349. <https://doi.org/10.1076/sesi.14.3.321.15842>
- Weick, K. E. (1995). *Sensemaking in organizations*. SAGE Publications.
- Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, *16*(4), 409–421. <https://doi.org/10.1287/orsc.1050.0133>
- Weiner, B. J. (2009). A theory of organizational readiness for change. *Implementation Science*, *4*(1), 67. <https://doi.org/10.1186/1748-5908-4-67>
- Woods, P., & Jeffrey, B. (1998). Choosing Positions: Living the Contradictions of OFSTED. *British Journal of Sociology of Education*, *19*(4), 547–570. <https://doi.org/10.1080/0142569980190406>

Paper 3: The Effect of Inspection on School Improvement Planning: Evidence from a U.S. District

Introduction

Prioritization of educational reform areas is an issue of national concern. Districts in the United States overwhelmingly rely on test-based accountability to promote school improvement (Figlio & Loeb, 2011; Hanushek & Raymond, 2005). Schools are incentivized to raise student achievement on standardized tests (e.g. Ladd & Figlio, 2008), yet test scores alone offer limited insight into specific reforms that might benefit a given school (e.g. Gagnon & Schneider, 2019).

An alternative approach to accountability is school inspection, which is widely used outside of the United States. Such an approach, using on-site evaluations, allows for a deeper assessment of schools. In this way, inspection feedback can guide planning and implementation of reforms (Ehren et al., 2013; Jones & Tymms, 2014). Gains in standardized test scores have been used to evaluate the effectiveness of school inspection (e.g. Allen & Burgess, 2012; Ehren & Shackleton, 2016; Hussain, 2015; Luginbuhl et al., 2009). Yet this offers limited insight regarding the influence of inspection on school planning and reform. In contrast, effectiveness can be evaluated focusing on whether inspection feedback leads to school reforms. This approach considers if comprehensive on-site evaluations can inform principals' actions. Further, it enables to see what areas of inspection are the most influential in promoting school reforms.

A crucial step prior to implementation of reforms is the school planning process (Matthews & Sammons, 2004). At this stage, school leaders prioritize areas for improvement and set strategic goals. Despite the wide use of inspection globally, no empirical evidence exists regarding its effect on school planning documents.

Prior studies have investigated the influence of school inspection on reform implementation European countries (e.g. Altrichter & Kemethofer, 2015; Dederling & Müller, 2011; Ehren &

Visscher, 2008; Gray & Wilcox, 1995; Ouston et al., 1997). This research is limited to post-inspection data that is self-reported. Interviews and surveys capture changes in school reforms based on inspection feedback. These studies are unable to capture causal effects of inspection since observations prior to inspection are not included. No prior research utilizes longitudinal data to evaluate the effect of inspection on school reform or improvement planning (Dedering & Müller, 2011). To address this gap, this study examines whether inspection feedback shapes school planning. Overall this study asks: Does school inspection feedback lead to changes in school planning? More specifically, my research addresses the following research questions:

- Do principals find school inspection useful for planning purposes?
- What areas covered by inspections influence principals to plan and/or implement changes in schools?
- What is the causal impact of inspections on the presence of key influential areas on School Improvement Plans (SIPs)?

To examine these questions, this study is based in one of the few U.S. districts to experiment with school inspection. This urban school district has conducted inspections for nearly a decade. A mixed method approach is used. In-depth interviews with school principals examine whether and how inspection has been useful for planning purposes and identify which inspection topics were more influential to inform school reforms. Then, a content analysis is used to establish the incidence of these topics on school improvement plans. Finally, a difference-in-differences approach is used to determine how inspection shapes the emphasis on these topics in school improvement plans. This is the first study to measure the causal effect of school inspection on school planning. In addition, this study provides empirical evidence in the U.S. of the potential of school inspection to inform school planning reforms beyond standardized tests.

Literature Review

School Change based on Inspection Feedback

Few empirical studies address the influence of inspection on school planning. Prior work indicates that schools tend to implement reforms after inspection (Cuckle et al., 1998; Dederling & Müller, 2011; Ehren & Visscher, 2008; J. Gray & Wilcox, 1995; Ouston et al., 1997; Verhaeghe et al., 2010). However, results are not consistent across studies. Considerable variation exists regarding the extent to which inspection recommendations lead to implementation of improvements (de Wolf & Janssens, 2007). For example, a study in Germany found that inspection led to an increase in reforms in a majority of schools (Dederling & Müller, 2011). Yet, a U.K. study found that only a small portion of inspection recommendations were implemented (J. Gray & Wilcox, 1995).

Prior research has found that how feedback is delivered can influence whether improvements are implemented (e.g. Ehren & Visscher, 2008; Gustafsson et al., 2015; Matthews & Sammons, 2004; Ouston et al., 1997; Penninckx et al., 2015). Greater implementation occurs when feedback is clear and explicit (Matthews & Sammons, 2004), shows school weaknesses (Ouston et al., 1997; Penninckx et al., 2015), and when shared goals are established between schools and inspectors (Ehren & Visscher, 2008; Ouston et al., 1997).

Reforms are also influenced by accountability pressure. There is evidence that principals who feel greater accountability pressure tend to be more attentive to inspectors' expectations and more responsive in terms of improvement actions (Altrichter & Kemethofer, 2015). Yet, pressure that is viewed as ill-intentioned might be detrimental to reforms. Implementation of reforms is less likely if inspection uses coercive methods (Gustafsson et al., 2015) or if the process is perceived as threatening (Visscher & Coe, 2003). Similarly, it was found that differentiated inspection models—where low performance schools have more intensive inspection—tend to be more effective to enabling reforms (Ehren, Gustafsson, et al., 2015).

A separate body of literature analyzes the causal effect of inspection on student achievement. This literature is also thin and far from conclusive. Most of these studies have found small positive effects on student achievement (e.g. Allen & Burgess, 2012; Ehren & Shackleton, 2016; Hussain, 2015; Klerks, 2012; Luginbuhl et al., 2009; Shaw, Newton, Aitkin, & Darnell, 2003). Yet, others have found no significant impact (e.g. Rosenthal, 2004).

The two bodies of literature described present a contrast. First, the existing literature on school responses to inspection, are based only on post-inspection observations and therefore do not estimate a causal effect of school inspection on developmental actions (de Wolf & Janssens, 2007; Dederling & Müller, 2011). Most of these studies focus on acceptance of inspection feedback and do not evaluate reforms within specific areas (Altrichter & Kemethofer, 2015; Ehren & Visscher, 2008; J. Gray & Wilcox, 1995; Ouston et al., 1997). Second, the literature on the causal effect of school inspection on student achievement does not address what aspects of inspection are responsible for gains. This study fills a gap and evaluates the causal effect of school inspection feedback on school improvement planning, identifying the effect on influential areas.

The Uses of School Improvement Plans

SIPs are strategic management instruments used as road map for school improvement. Although SIP content varies across districts, considerable similarities have been found (Mintrop et al., 2001). Common content areas in SIPs include: 1) establishing a vision, 2) assessing needs, 3) setting strategic goals and actions, and 4) using measurable performance metrics to evaluate past performance and monitor progress (Fernandez, 2011; Redding & Searby, 2020; Strunk et al., 2016). A key function of SIPs is establishing school priorities. In the management literature, this has been associated with better organizational performance (Chun & Rainey, 2005; Hines, 2017). Regarding the effectiveness of planning documents in education, the empirical literature is very thin. Fernandez (2011) found a strong association between the quality of school planning and student performance in reading and math.

The SIPs have been a central instrument for high-stakes accountability systems in the United States (Mintrop et al., 2001). The New Public Management reform since 1980s promoted the use of strategic planning in public agencies. This management technique used by successful corporations was viewed as a practice that would enable rational planning and greater efficiency in the public sector (Berry & Wechsler, 1995). Improvement plans emphasize performance quantification and promote accountability. This aligned well with the emphasis of test-based accountability on data driven decision making and a focus on outcomes (Fernandez, 2011; Redding & Searby, 2020).

SIPs have played a central role in U.S. education reforms (Armstrong, 1982; Doud, 1995; Fernandez, 2011; Mintrop et al., 2001; Strunk et al., 2016). Federal and state mandates have widened their use. At the national level, the Elementary and Secondary Education Act (1965), No Child Left Behind (2001), and Every Student Succeeds Act (2015) have successively advanced requirements of submitting SIPs to state education agencies in order for low performing schools to access federal funds (Meyers & VanGronigen, 2019; Mintrop et al., 2001). In addition, states agencies often provide rules, guidelines, and templates for developing these plans.

SIPs have been used as a management tool to align the goals of accountability systems and individual schools. This has resulted in widespread internalization of state goals into the operations of individual schools. For example, a content analysis of SIPs from low-performing schools in three states found relatively uniform goals and activities, as schools adopted goals mandated by state agencies (Mintrop et al., 2001). While filing SIPs is mandatory, this does not necessarily mean that they will be a key planning tool for schools (Cuckle et al., 1998; Meyers & VanGronigen, 2019). One study found that the use of inflexible bureaucratic practices resulted in 80% of principals adopted “satisficing behavior” (i.e. “good enough” practices) for SIPs; this included resubmit previous years’ plans or solely focusing goals on test scores (Meyers & VanGronigen, 2019).

District Background

This research focuses on a large urban school district in the United States that has used inspection as a supplemental mechanism for school accountability and improvement during the last 10 years. Like other school districts in the United States, my case study relies primarily on high-stake testing for accountability purposes. The main accountability instrument is the Performance Framework, which rate schools based on a variety of performance indicators. Yet, standardized test results are the most influential rating component. These ratings guide incentives, sanctions, and support actions for schools.

School Inspections

The district has used school inspection since 2012. Inspections are focused on low performing schools, based on the Performance Framework. However, the district has discretion in selecting schools for inspection. This study focuses on inspections conducted in the school years 2016-17 and 2017-18. These are the two years with the most school inspections. The procedures used for inspection remained the same during these school years; the process was changed for visits in 2018-19 and later.

School inspections are facilitated by a contractor. Teams of 3 to 4 inspectors make school visits; each team includes two contracting staff and at least one representative from the district department of education. A protocol guides the process. Inspections entail a two-day, on-site evaluation of school quality. During this visit, inspector review school documents, observe classrooms, and conduct interviews and focus groups with administrators, teachers, parents and students.

The scope of topics covered by the inspection is broad, yet, it has an instructional lens. Areas covered by the inspection include classroom instruction, support to students, professional development, school climate and culture, leadership, and relationship with the families and community. The process is highly structured. Inspectors use rubrics to evaluate classroom

observations and questionnaires to conduct interviews and focus groups. At the completion of the visit, the inspection team meets with school administrators to provide feedback on findings and discuss improvement strategies. After the inspection, a written report summarizing the findings is sent to the school.

The final report includes an evaluation of each domain (e.g. “Instruction”), a rating of each quality criteria (e.g. for the criteria “Classroom instruction is intentional, engaging, and challenging for all students,” the school “does not meet,” “partially meets,” “meets,” or “exceeds” expectations) and an evaluation of the topics within each quality criteria (e.g. “Instruction does not require students to use and develop higher order thinking skills”). The report highlights the school’s “strengths” and “areas of growth.” Finally, based on the discussion among the inspection team and school leadership team, the document reports areas to prioritize, goals set, and measures to evaluate success. There are not follow up instance after inspections.

School Improvement Plans

The district follows the state requirement for all schools to submit a SIP every year. High performing schools can request to only submit plans every two years. The SIPs build on the strategic planning of the schools, looking for a consistent format to capture planning efforts and aligning to state and federal requirements for multiple programs and grants. The State provides a template as well as detailed guidelines and assistance for developing the plans. In addition, the SIP intentionally looks to provide enough flexibility so that the planning process is meaningful for the schools.

A major part of the SIP is a narrative section, which is unstructured. Schools describe their plan in detail, giving the opportunity to build a coherent case that summarizes the overall plan. This narrative typically includes a description of the school, mission and vision, climate and culture, instructional models, family and community engagement, leadership and staff, diagnosis of challenges and performance problems, current activities, programs, and partners, past support and

grants, a deep analysis of prioritization of areas of focus, strategies for improvement, and future plans.

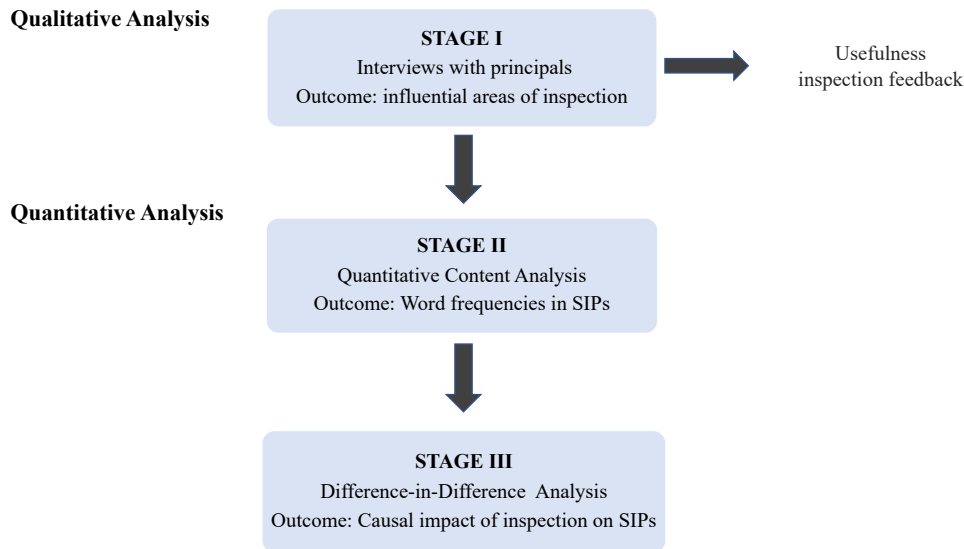
All other SIP sections are highly structured and rely on performance indicators. These sections include trend analyses, performance challenges, root causes of performance challenges, prioritized areas, improvement strategies, action plans, and monitoring the impact and progress of the action plan.

This study focuses on the narrative section of the SIPs. Analyzing this section, as opposed to the whole plan, enables me to focus on the main message school leaders decide to highlight. Furthermore, it avoids the repetition of topics across sections. Its unstructured format facilitates content analysis, which is used to capture the message of this type of texts and avoid focusing on the topics set by the plans' templates.

Research Design

This study uses mixed methods with a quantitatively driven design (Morse, 2010). The research design is sequential (Greene et al., 1989; Teddlie & Tashakkori, 2009) and comprises three stages (Figure 2). The first stage consists of in-depth interviews with principals of inspected schools, looking to identify their perceptions regarding usefulness of inspection and influential areas (i.e. areas of reform that were planned or implemented based on feedback). Based on interview responses, most influential areas are identified. The second stage uses content analysis to measure the presence of these topics in the SIPs. Quantitative content analysis involved coding text into categories and counting the frequencies of occurrences within each category (Ahuvia, 2001), which are used as a proxy of topic importance. This is an intermediary step for the following stage. Lastly, a difference-in-differences analysis tests for evidence of a causal effect of inspections on SIP improvement areas. I analyze if there is a change in focus within SIPs of inspected schools, compared to schools not inspected.

Figure 2. Research Design



Stage I. Interviews with School Principals

In-depth semi-structured interviews with principals from schools that were inspected in school years 2016-17 to 2018-19 were conducted. The goal of the interviews is to evaluate how inspection feedback was useful and what areas covered by inspection were more influential leading to changes in the schools. A total of 55 schools were inspected at least one time during this period (44 schools were inspected one time, 10 schools twice, and 1 school three times). All 55 principals were invited to participate in the interviews; 16 were interviewed. Participants were informed that interview responses would be anonymous and would be attributable to a pseudonym. They received US\$ 25 gift card after participation.

Interview questions inquired about the perceived usefulness of inspection feedback, main ongoing or recent changes implemented in schools, what motivated the changes, and what changes were based on inspection feedback. To assess responses, a code book was developed inductively (see codebook in Appendix A). Codes of usefulness inquire on the ways in which the inspection were

useful for planning purposes, such as narrowing the improvement focus, reaffirming existing goals, or gaining legitimacy. Codes for changes implemented inquire on the different sources that led to changes in the schools, such as “principal initiative,” “data analysis,” or “inspection feedback.”

To ensure reliability, an independent-coder approach was used. First, interview transcripts were independently coded by two researchers. Then, codes were compared for agreement. An iterative process was followed until reaching at least 80% agreement, based on a pooled Kohen Kappa indicator (De Vries et al., 2008). We achieved an 83% of agreement. Next, I used individual codes to search for themes and patterns both within individual interviews and across interviews. I identified the presence and absence of codes based on frequency charts and cross-tabulations.

Influential areas based on inspection feedback

Changes implemented or planned based on “Inspection feedback” are coded based on areas of change. These areas stem from analyzing the inspection protocols. Eight areas of inspection were identified: 1) Community Involvement, 2) Climate & Culture conducive to Learning, 3) Instructional Practices, 4) Leadership, 5) Professional Development, 6) Support to Students, 7) Teachers-Administrators Collaboration, and 8) Other Organizational Issues.

Interviews reveal that most principals (88%) implemented or planned changes based on inspection feedback. About 80% of principals focused on changes related to *Instructional Practices* and/or *Climate & Culture conducive to Learning* (from now on “Climate & Culture”). In most cases, several areas of change are addressed simultaneously. About 70% of principals who mention reforms in these two areas also implemented changes in other areas.

Instructional Practices is the most commonly mentioned reform area that principals address as a result of inspection feedback (see Table 6). This is not surprising, since instruction is at the center of the inspection process. Instructional topics mentioned repeatedly by principals include promoting higher order thinking, depth of questioning in classroom, setting clear expectations, and improving formative assessments. This improvement area is closely followed by *Climate & Culture*,

which half of principals address in implemented and planned reforms. The scope of topics was wide, including “how the school culture impact student learning,” attitudes and expectations toward students, behavioral interventions, and social emotional learning.

Table 6. Influential Areas - Changes Implemented in Schools based on Inspection Feedback

Area	Number of Principals	Percentage
Instructional Practices	9	56%
Climate & Culture conducive to Learning	8	50%
Other Organizational Issues	5	31%
Leadership	4	25%
Professional Development	3	19%
Support to Students	3	19%
Staff Collaboration	3	19%
Community Involvement	-	-

Neither of the other specific categories were mentioned by more than four principals. The category *Other Organizational Issues* (31%) includes changes on “teacher schedules”, “building a strong vision” or “data meaning.” Some principals mentioned changes that appear to be too broad or vague, such as “setting systems in place,” and “consistency in structures.”

To capture the influence of school inspection and to limit overlap across categories, the research focuses in the two areas were most principals implemented changes: *Instructional Practices* and *Climate & Culture*. These identified areas were then used to develop categories for the content analysis and the difference-in-difference analysis in the following stages.

Stage II. Content Analysis

A dictionary-based, quantitative content analysis (Krippendorff, 2013; Riffe et al., 2014) of SIPs identifies the presence of words associated with influential areas emphasized by school principals. In order to evaluate the impact of inspections on school planning during the school years

2016-17 and 2017-18, the analysis includes 399 improvement plans from all 205 K-12 public schools in the district.

The areas of *Instructional Practices* and *Climate & Culture* are the categories of the content analysis. To define the scope of these areas I analyze the inspection protocols, leading to the following definition for inclusion for the two areas:

- i. **Instructional Practices.** Focus on high quality instructional practices and interactions. Purposeful, intentional, and engaging teaching. Emphasis on rigor and higher order thinking skills. Group work and cooperation. Feedback to students and ongoing assessments. Alignment with the common core state standards.
- ii. **Climate & Culture.** General school culture and climate conducive to learning. High behavioral and academic expectations. Rewards for positive behavior and consequences for misbehavior. Clear expectations, respect for school norms. Consideration for the whole child and support for emotional learning. Supportive, collaborative, and caring interactions with the students.

The content analysis uses a semi-automatic dictionary building process (Brier & Hopp, 2011; Deng et al., 2019; Grimmer & Stewart, 2013; Neuendorf, 2017). All phrases and words included in the dictionary stem from the SIPs. The dictionary building process was conducted using WordStat 8 software, following these steps in an iterative process:

- 1) Corpus creation: All narrative sections of the SIPs for school years 2016-17 and 2018-19 conform the corpus of analysis.
- 2) Initial word frequency list: A frequency list with all the words included in the documents is created. I identify 8,309 unique words.
- 3) Pre-processing: Removes all “stop words” or function words that do not convey meaning, such as conjunctions and prepositions. I also exclude words and phrases that appear in less than 5% of the documents (i.e. less than 20 SIPs). This results in 1,201 words.

- 4) Initial phrase frequency list: A frequency list with all the phrases—at least 2 words together—included in the documents is created. I identify unique 398 phrases.
- 5) Entry identification and classification: Words and phrases are the basic unit for classification. A selection of all phrases and individual words was conducted manually, eliminating those that clearly do not belong to the category analysis. The process started with most frequent phrases which are usually more “context resistant” (Conway, 2006; Deng et al., 2019). This results 411 phrases and words.
- 6) Consolidation: Words included in the preliminary dictionary were further reduced through word stemming (i.e. the stem emotion* counts the words “emotions,” “emotional,” and “emotionally”). Alternative spelling and acronyms were added as substitute of dictionary words and phrases. This results in 362 phrases, words, word stems, and acronyms.
- 7) Contextual validation: A key-words-in context analysis was conducted. I assessed each term in context—reading the whole sentence in which the terms are used—to decide whether it belong or it does not belong to the category of analysis. If less than 50% does not belong, I exclude the word from the dictionary. If 50% to 80% belongs to the category of analysis, further analysis was conducted. This included checking other word forms and co-occurrence with other words to attain more precision. This refining process was conducted until over 80% of key-words-in context rendered true positive results; at this point, the word is considered to belong to the main categories of analysis (Bengston & Xu, 1995; Deng et al., 2019).

This step led to directly including of 52 terms (>80% of true positive), directly excluding of 246 terms (<50% of true positives), and further considering of 64 terms.

Partial result after iterations: 104 terms.
- 8) Extensions: Extension of the dictionary considering word co-occurrence, synonyms and antonyms of pre-selected words, to include words that might have been overlook or excluded

due to low frequency. A misspelling identification analysis was conducted to detect false negatives. Then, validation of key-words-in context, described above, was repeated.

Final results: 119 terms.

An 80% cut-off criteria used to determine word inclusion or exclusion in the dictionary is based on precedent from prior literature (Bengston & Xu, 1995). Any threshold risks dropping potentially relevant entries (Deng et al., 2019). In addition, semi-automatic dictionary building process will inevitably include some categorization errors. To assess the overall validity of the dictionary, I compare software and human coding for a random sample of 10% of the SIP documents. I find a Scott Pi of 79%, indicating a good level of agreement. The outcome from this stage is a frequency of words within the categories of analysis—most influential areas—for each SIP.

Word Frequencies in SIPs - Most Influential Areas

The dictionary includes 119 terms—including phrases, words, word-stems, and acronyms: 62 terms for Instructional practices and 57 terms for Culture & Climate (See the dictionary in Appendix B). For the classification of Instructional Practices, the categorization focuses on transversal issues across subject. In contrast, it omits subject specific issues— such as “literacy”, “phonics”, or “manipulatives”—; it also omits references to standardized tests, performance framework, and specific support and curricular programs. For example, the dictionary includes terms such as “checking for understanding,” “grouping,” and “formative”; but excludes “SAT,” “school performance”, or the program “Amplify Science.” For the classification of Climate & Culture terms, the dictionary excludes professional staff aiming to provide support to students—such as “psychologist” or “social worker”—.

The dictionary was designed to make the main categories mutually exclusive (Riffe et al., 2014). Terms that overlap in the two main categories, such as “student engagement,” were classified

as *Climate & Culture*. Since the whole inspection process has an instructional lens, this decision was made to increase the chance of capturing climate and cultural issues.

The narrative section of the SIPs provides flexibility to principals to address a wide range of issues. To provide a sense of the scope of topics, clouds of the most frequent terms covered in SIPs for the school years 2016-17 and 2018-19 were created (See Appendix C). Some of the most frequent topics include leadership, English language learners (i.e. “English language”), professional development, demographics (i.e. “reduced lunch”), and performance (i.e. “student achievement”).

Table 7 present the content analysis coverage and the frequency of the main categories. Overall, the dictionary words cover close 2% of the total number of words of the total words for both years (See Table 7). Nonetheless, these topics cover 12% of the total number of sentences in 2016-17 and 19% for 2018-19. The word frequency on the *Instructional Practices* category remained stable over the two years. In contrast, the frequency of the *Climate and Culture* category grew more than doubled. The frequency list of key terms for each SIP is the input for the statistical analysis.

Table 7. Content Analysis Coverage and Term Frequency

	School Year	
	2016-17	2018-19
Dictionary Coverage		
% words	2.0%	2.1%
% of sentences	12.0%	18.7%
Category Frequency		
Instructional Practices	612	649
Climate and Culture	600	1278
N	196	203

Note: % of SIP words excludes “stop words.”

Stage III - Statistical Analysis

I use a difference-in-differences analysis to evaluate the causal impact of inspection on the presence of key influential areas on the SIPs. The outcome measure stems from the content analysis

in the areas of 1) Instruction and 2) Culture & Climate, as a proxy of attention. The “word count” is the sum of both categories.

The analysis focuses on the impact of inspections conducted in school years 2016-17 and 2017-18. In a given year, the SIPs that guide school planning were prepared at the end of the previous school year. For example, the SIPs the guide schools in year 2016-17 were prepared in late spring or summer of 2016. SIPs that reflect the pre-intervention period in my study are those that guide schools in 2016-17; these are prepared before the inspections occur. Post-intervention SIPs are those from 2018-19; these were written after inspections.

I include schools that have SIPs available for both the year reflecting pre-intervention plans (2016-17) and post-intervention plans (2018-19). I exclude schools with inspections within two years prior to the study period; this excludes 26 schools with inspections in 2014-15 and/or 2015-16. My final sample comprises 160 public schools (79% of public schools in the district). My treatment group has 31 schools with at least one inspection in 2016-17 or 2017-18. My comparison group has 129 schools that did not have inspections and serve as controls.

Table 8 presents summary statistics at the school level for inspected and not inspected schools in the periods before and after inspection.³ On average, in the post-treatment period inspected schools have a significant higher word count on influential topics, compared to not inspected schools ($p < .001$). Regarding school characteristics, on average, inspected schools have a higher proportion of low-income students, as indicated by the portion of student receiving free and reduced-price lunch. No other observed characteristic differs significantly between treatment groups. Student achievement, measured by the state standardized tests is also considered. Tests scores are unavailable for 31 schools in the sample. Average test scores in English Language Arts and Math are presented in

³ School demographic and school characteristic data were obtained from the state Department of Education website. Since the School District required not to be named, the name of the State is not disclosed either to avoid facilitating the district identification.

Appendix D. Schools not inspected have higher test scores, on average, in both subject areas and in both years ($p < .001$).⁴

Table 8. Summary Statistics

Outcome Variable	School Year 2016-17		School Year 2018-19	
	Not Inspected	Inspected	Not Inspected	Inspected
	Mean	Mean	Mean	Mean
Word count	6.02 (7.84)	6.29 (6.68)	7.6 (7.02)	13.45*** (5.37)
School Characteristics				
Enrollment	458.93 (310.31)	524.19 (274.32)	471.81 (328.16)	495.90 (248.36)
FRL (%)	0.66 (0.28)	0.77** (0.20)	0.66 (0.28)	0.75** (0.19)
Black (%)	0.13 (0.12)	0.14 (0.16)	0.13 (0.12)	0.14 (0.16)
White (%)	0.24 (0.25)	0.19 (0.17)	0.25 (0.25)	0.18 (0.17)
Hispanic (%)	0.56 (0.29)	0.59 (0.25)	0.55 (0.29)	0.59 (0.25)
n	129	31	129	31

Notes: Standard deviations are in parentheses. FRL = free and reduced-price lunch.

** : significant difference at the 5% level; difference between inspected and not inspected schools in the indicated school year, based on two-sample t-tests .

*** : significant difference at the 1% level ; difference between inspected and not inspected schools in the indicated school year, based on two-sample t-tests

The parametric DD analysis aims to capture the causal impact of inspection on the presence key topics in the SIPs, controlling for covariates and school fixed effects. The outcome variable (word count) distribution shows a positive skew, more pronounced in 2016 (skewedness =3.2) than 2018 (skewedness =0.9). To address skewness, rather than use a canonical DD model, I considered

⁴ Only 128 of the panel schools have test score data available.

three alternatives: a log-linear model, a Multilevel Mixed-effects Generalized Linear Model with a negative binomial distribution, and a Multilevel Mixed-effects Generalized Linear Model with a Poisson distribution. The log-linear model was chosen given its better fit, minimizing the error dispersion.

My DD model is specified as:

$$\ln(Y_{it}) = \beta_0 + \beta_1 Post_t + \beta_2 Insp_i + \beta_3 Insp_i * Post_t + \theta X_{it} + \delta_i + \varepsilon_{it}$$

where Y_{it} represents the key word count within a SIP for school i in year t , $Post_t$ indicates the post-inspection time period (school year 2018-19), $Insp_i$ indicate that the school had an inspection, X_{it} is a vector of school-level characteristics, δ_i are the school fixed effects, and ε_{it} is the random error term. $Insp_i * Post_t$ is an interaction term between inspected schools and an indicator for the post-inspection period; β_3 is the DD treatment effect. Since this is a log-linear model, regression estimates are calculated in log points. A transformation is needed to interpret results as percentual points. For example, β_3 , the DD effect, has the following interpretation: on average, inspected schools have $(e^{\beta_3} - 1) * 100$ percentage points increase/decrease in key words, in comparison to non-inspected schools in the post treatment period.

The school fixed effects capture the effect of unobserved, time invariant factors that might influence the outcome, such as principals' ability to plan. School level characteristics include number of students enrolled, percent of student receiving free and reduced-price lunch, race, and ethnicity (% of white, black, and Hispanic students).

Results

Interview Analysis - Perceived Usefulness of Inspections

Most principals find that inspection is useful for planning purposes. Responses are grouped in three non-mutually exclusive categories that indicate that inspection allowed principals to: 1) prioritize the school focus of planned improvement (80% of principals), 2) confirm their prior

diagnosis or existing goals (75%), and 3) increase legitimacy among school staff and with the school district to implement changes (50%). Those principals who highlight how inspection brought legitimacy within the school to implement changes, explain that inspection facilitated a collective process of planning. Only three principals out of the sixteen interviewed did not find inspections useful. Overall, interview responses demonstrate that the inspection process supports planning, and this planning goes beyond the areas identified in the district's Performance Framework, which is heavily focused on standardized test results.

Improved Prioritization. The inspection feedback helped 80% of principals plan strategically, through improved prioritization of the school focus. This included identifying new improvement areas and narrowing the current focus. Many principals highlight the wide variety of issues covered by inspection, allows for more comprehensive reforms.

Several principals note that the inspection provided an *opportunity* to make changes that their schools "needed." Principal Tyler explains that the inspection not only uncovered areas that they needed to work on, but also gave them the "the time and space to actually work" and "restructure the strategic planning." More concretely Principal Linda recounts that the inspectors consistently saw disciplinary issues in the classroom and lack of systems to deal with it, which led "throughout the school" to ask questions, such as what are they "doing wrong," and what should they change. She sees these questions as an opportunity to define their focus as well as action steps.

Facilitating *understanding* of the problems and providing evidence were other aspects that many principals highlight as useful. Principal Sebastian reflects that the inspection informed his "understanding as a principal, of some deeper issues inside of the school." Principal Thomas explains how he selected a focus area for improvement after the inspection illuminated a specific challenge and provided evidence:

[After receiving the inspection feedback] we were able to work as a team and have things broken down in such a way that people felt like it was something that we could and needed to

focus on. So I think that the whole process just made it really tangible for us to have clear things to focus on for us to then say, "Okay. Rigor is the one that we keep hearing and seeing, and so that's the one that we're going to take as our next step as a school to really look to move forward with."

Most principals used inspection feedback to prioritize and define their focus in improvement plans. Principal Mark explains that there are many worthy areas he “could choose to tackle and [the inspections] really kind of helped ... hone in on two areas for my major improvement strategies for my school improvement plan. They really were crafted around that.” In addition, many principals provide examples of how the inspection was useful for advancing to implementation. For example, Principal Mary explains that the feedback helped them “to start writing out some action steps based on the highest leveraged area that the school could focus on.”

Diagnosis Confirmation & Goal Reaffirmation. About 75% of principals state that the inspection feedback was useful in confirming their diagnosis or reaffirming existing goals. Feedback often confirmed what they already suspected and served to validate that they were “headed in the right direction” (Principal Amy), “doing lots of really good things” (Principal Thomas), or “working on the right things” (Principal Tyler). This confirmation was useful for Principal Nicholas in justifying their current focus:

Before the [(inspection)] year, we had included Social-Emotional Learning as a major improvement strategy for our improvement plan. [The inspection] ... helped confirm that, that was a valid area of focus, to invest in. So, where we might've felt tempted to just do away with it, now there's just a lot more excitement to continue with it and to keep in the actual improvement plan that we submit to the state. The fact that it was called out so explicitly in the [inspection] was pretty surprising.

Inspection also confirmed what areas were problematic, as explained by, Principal David:

... what the inspection did help do is provide more clarity and specifics around things that maybe I thought were gaps... And I think that provided opportunities for me to kind of get back the dots.

Gained Legitimacy. Half of the principals found that the inspection brought legitimacy with the staff and the school district the school to implement changes in the strategic planning. Inspections accomplished this through three main pathways: 1) providing evidence to justify selected focus areas for improvement planning; 2) incorporating the views of school staff.; 3) establishing a common ground for planning among the school staff.

Many principals agree in their view that the inspections legitimized their improvement areas and strategies. In most cases, gained legitimacy seems to play a more relevant role within the school, as illustrated by principals Sarah and David:

Principal Sarah: we talked about instructional rigor, and that was one that we really latched onto... it was useful to show the data to our teachers, because our teachers felt that they were very rigorous in their instruction, and for us to go back, and say, "Here's this piece of [inspection], that's actually not the case." Because sometimes when you say something, people don't always believe it, but when you have the data behind it, it really hammers home in a different way. Then, for us to say, "Our school-wide focus, we're going to focus on rigorous instruction," that helped us out.

Principal David: ...the biggest place where the inspection was useful was for me to be able to say no. We as a leadership team have seen this gap, and it's confirmed by this outside source. And it's confirmed by our data. This is something we need to address because clearly what we're doing is not working.

The inclusion of inputs from a variety of school stakeholders during the inspection visit, were used as an additional source of legitimacy to promote changes in the schools. This is the case made by principals Matthew:

I knew that we were falling short on some of our work on observation feedback in terms of what I wanted. What I didn't understand was the teachers were also wanting it ... And then..., you're honoring teacher voice, ... so you're able to say ... "Remember when you guys all shared with the folks that you wanted more observation feedback, well I've created a teacher lead position that's going to help with that."

Half of the principals found that the inspection facilitated setting a *common ground* with the staff for planning and implementing changes. The process was useful to check if the staff follow and understand instructional processes (Principal Monica), to “start conversations about institutional practices” (Principal Sarah), and to establish a “common understanding” (Principal Mary).

Finally, for some principals, the inspection report was also perceived as a source of legitimacy for the district. Principal Nicholas makes the case:

We've been investing in areas that have traditionally not been supported outside of the school or by the district, because the push has always been academic. And so, by focusing on these other areas on the whole child, on Social-Emotional Learning, on the school culture, is a big risk that we would not be accepted as good leaders. So, it's very helpful ... to bring more support to us, to encourage District leadership to be more supportive of our effort.

One size does not fit all. Although a majority of principals (81%) found the inspection process useful for planning purposes, three principals did not. Reasons for these stances include that inspections cannot uncover the “true culture” (Prin. Brian) of a school in a short visit, “unnecessary” (Prin. Ashely) because school administrators could do the evaluation themselves, or not informative enough (Prin. Paul). In addition, other principals mentioned specific aspects of the feedback they found unhelpful. The most common reason was the timing to implement changes or the fact that they already had reforms in place (Principals Monica, Tyler, Sarah, and David). Two principals also find that the quantity of information was “overwhelming” (Prin. Nicholas) and not helpful enough providing “actionable next steps” (Prin. Sebastian).

Difference-in-Differences Analysis

The DD analysis demonstrates an increased use of key words in SIPs that is significant among inspected schools. First, I calculate the non-parametric difference-in-difference value (Table 9). On average, inspected schools have 5.6 more words in their SIPs in the post-inspection period related to Instructional Practices and Culture & Climate in the post-period. To put the quantity of words in context, in school year 2018-19, on average 21% of the sentences were classified as *Instructional Practices or Climate and Culture* (Table 10), which contains the 13.5 key words (Table 9).

Table 9. Non-parametric Difference-in-Difference

	School Year	Word Count Mean	First Difference	Diff-in-Diff
Inspected schools (N=31)	2016-17	6.3	7.2	5.6
	2018-19	13.5		
Not Inspected Schools (N=129)	2016-17	6.0	1.6	
	2018-19	7.6		

Notes: The *first difference* is the average word count of 2018-17, less the word count for 2016-17, for each group. The difference in differences, is the *first difference* of Inspected schools minus Not Inspected Schools.

Table 10. Content Analysis Coverage for Panel

Treatment	School year	% SIPs words	% SIPs sentences
Inspected n= 31	2016-17	1.9%	11.4%
	2018-19	2.6%	21.1%
Not inspected n= 129	2016-17	2.0%	12.2%
	2018-19	2.0%	17.5%

Note: % of SIP words is average keyword count over the total words, excluding “stop words,” in the SIPs; % SIP sentences is the proportion of sentences including keyword over all sentences in the SIPs.

Table 11 presents the results of the log-linear DD model with five different specifications: (1) the base model without school fixed effects nor covariates (Model 1), (2) the base model with selected school demographics (Model 2), (3) model 2 with test scores (Model 3), (4) model 3 with school-level fixed effects (Model 4), and (5) model 2 with school-level fixed effects (Model 5). The DD estimates are statistically significant in all five models.

Model 3 and Model 4, which include test results, do not include the 31 schools that lack test scores; this affects the treatment group (5 schools) and comparison group (26 schools). In model 5, the preferred model, the DD estimate is 0.666 log points ($p < .01$), which indicates that inspection, on average, results in a 94.6 percentage point increase in key words.⁵ These results show that inspection have a significant impact on the focus on schools planning.

Comparing model (1) with models (2) and (5), shows that the basic model is sensitive to adding covariates and school fixed effects. By including these controls, the magnitude of the DD estimate is reduced. In all the models, the DD is statistically significant ($p < .01$). Model (2) indicate that larger school have a lower effect than smaller schools, which indicate that school inspection might be more effective in informing and redirecting school planning in smaller schools. The significance disappears when school fixed effects are added in model (5). The DD estimate sensitivity to school fixed effects, indicate that there are unobservable factors within the school that influence the results (e.g. principals' ability to plan). Overall, results appear to be robust. Coefficient estimates are consistent across model specifications – with varying covariates and sample composition.

⁵ This is calculated as: $(e^{.666} - 1) * 100 = 94.6$ percentage points

Table 11. DD Regression Results

	Word Count (Logged)				
	(1)	(2)	(3)	(4)	(5)
Post	0.319*** (0.090)	0.327*** (0.090)	0.414*** (0.098)	0.303*** (0.105)	0.326*** (0.088)
Inspected * Post	0.757*** (0.188)	0.731*** (0.186)	0.765*** (0.201)	0.533** (0.213)	0.666*** (0.186)
FRL (%)		0.171 (0.638)	-0.790 (1.015)	-7.238** (3.242)	-2.581 (1.999)
Logged Enrollment		-0.395*** (0.114)	-0.091 (0.122)	-1.067* (0.576)	-0.444 (0.378)
% Black		-0.493 (2.029)	0.801 (2.238)	-4.796 (5.320)	-3.148 (4.512)
% White		-0.290 (1.781)	0.348 (2.045)	-7.233 (4.560)	-2.769 (3.999)
% Hispanic		-0.113 (1.748)	1.026 (1.920)	0.848 (5.151)	0.382 (4.289)
English Language Art			-0.016* (0.010)	0.006 (0.017)	
Math			0.007 (0.010)	-0.005 (0.017)	
Constant	1.480***	3.917**	8.342	13.812	6.738
N	320	320	258	258	320
R-squared (within)	0.237	0.244	0.288	0.364	0.269
School Fixed Effects	No	No	No	Yes	Yes

Note: Robust standard errors in parentheses.

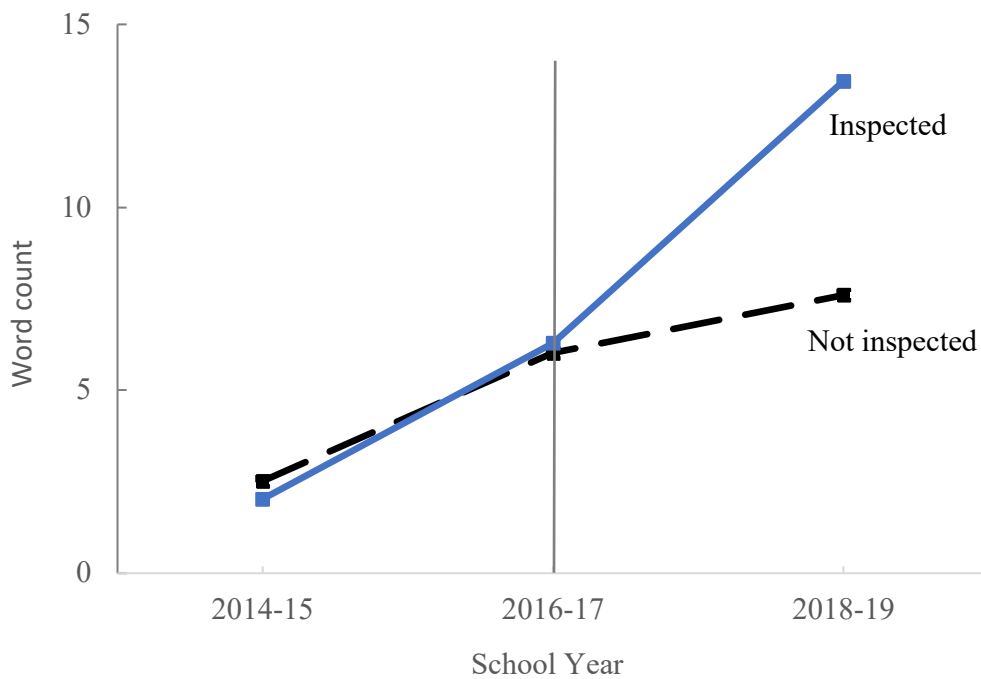
*** p<0.01, ** p<0.05, * p<0.1

A series of robustness checks were conducted to provide evidence in favor of the casual interpretation of the DD estimates were conducted: 1) graphical examination for parallel trends, 2) balance tests, and 3) placebo tests.

I visually examine the parallel trend assumption, which is critical to the validity of DD models. I compare the outcome variable Word Count in SIPs reflecting pre-inspection years (2014-15 and 2016-17) between inspected schools and those not inspected. Since high performing schools

can opt to submit a plan every two years, this robustness check relies on the school year 2014-15, rather than 2015-16. This alleviates concern that schools submit the same report in the pre- and post-inspection years. Reports prior to 2014-15 are excluded since a narrative section was not required in these early years. Figure 3 indicates that the average word count was similar for inspected and non-inspected schools, prior to inspections. This evidence provides support in favor of the parallel trend assumption underlying the DD model. In the post-inspection year (2018-19), inspected schools have a higher count of inspection-related words.

Figure 3. Parallel trends: Word Count for Inspected vs. Not Inspected Schools



Note: Word count represents the average number of keywords related to Instructional Practices and Climate & Culture in the SIPs (see Research Design section).

A balance test examines whether differences in attributes of inspected and not inspected schools are stable over time and that there is not association between the treatment exposure and the covariate distribution. The test uses a DD model take the covariates from the original model as outcome variables. Table 12 presents the results of this model; the DD coefficient estimate is not

significant for any of the outcomes considered: enrollment, percent FRL, percent black, white, and Hispanic, and test results. There is no evidence of attribute imbalances in my DD models.

Table 12. Balance Tests

VARIABLES	Logged Enrollment	% FRL	% Black	% White	% Hispanic	English Language Arts	Math
Inspected * 2018-19	-0.06 (0.05)	-0.01 (0.01)	-0.00 (0.01)	-0.00 (0.01)	0.01 (0.01)	1.20 (1.72)	-2.60 (1.92)
N	320	320	320	320	320	258	258
Number of Schools	160	160	160	160	160	129	129

Note: Robust standard errors in parentheses.
 *** p<0.01, ** p<0.05, * p<0.1

Lastly, I conduct two placebo tests to examine alternative explanations (Cook et al., 2002). The tests include, (i) a DD model using total word count of the SIPs as a dependent variable; and (ii) a DD model using the word count of words related to inspection, but not included in the set of words associated with Instructional Practices and Climate & Culture (see Appendix E for list of the alternative, inspection-related words). Apart from the change in the dependent variables, the placebo tests have the same specification as Model (5) of Table 11, this includes school fixed effects. Table 13 present the DD estimators of the two placebo tests. Results show a statistically insignificant DD estimator for the two placebo test models. This provides additional support for a causal interpretation of the DD estimate.

Table 13. Placebo Tests

	(1) All Words (logged)	(2) Key Words (logged)
Inspected * Post	0.172 (0.143)	0.024 (0.137)
N	320	320
R-squared	0.284	0.350
Number of Schools	160	160

Note: Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Overall, the DD estimates indicate that inspected schools nearly double their use of keywords related to instructional practices and climate & culture conducive to learning. All robustness checks—visual examination of parallel trends, balance tests, and placebo tests—support the causal interpretation of the difference-in-differences estimates. This provides evidence of the significant impact that inspection can have on school planning and selection of priority areas for improvements.

Conclusions

This study finds evidence that school inspection can influence school planning. Results indicate that inspection shifted the focus of planning documents. In addition, principals perceived inspection as useful for planning purposes. The value of this study is twofold. First, it is the first to assess the causal impact of inspection on school planning. Second, it provides empirical evidence in the United States of the effectiveness of inspection in influencing school reform, based on comprehensive onsite evaluations that go beyond standardized test scores.

Principals indicated in interviews that inspections were useful in informing reforms, both in terms of planning and implementation. Inspector feedback helped school leaders to prioritize improvement areas, reaffirm prior diagnoses and goals, and gain legitimacy with school staff and the district. These factors are relevant for planning. First, setting priorities is the primary function of

improvement plans and is associated with better organizational performance (Chun & Rainey, 2005; Hines, 2017). Second, obtaining evidence confirming principals' diagnosis and goals provides support to sustain long-term reforms (Armenakis & Harris, 2009). Finally, staff participation in the inspection process served to establish common ground. Through planning as a team, proposed reforms had greater legitimacy among staff.

The broad scope of inspection offered an opportunity to consider areas of reform not addressed by the district accountability framework, which is heavily based on standardized test results. Most principals decided to implement changes as a result of the inspection feedback. Interviews indicated that despite the wide array of schooling issues evaluated by inspections, 80% of principals decided to implement changes in two areas of inspection—instructional practices, climate and culture—.

The difference-in-differences analysis found that inspection led to measurable changes in school planning. Inspection shifted the focus of school planning documents. Inspectors' evaluations led to significant increase in text devoted to topics within the most influential areas—instructional practices and climate & culture conducive to learning. The amount of text devoted to a topic is used as a proxy of importance to school principals. Inspected schools devoted 11% of sentences in planning documents to these topics prior to inspection, which rose to 21% after the intervention. Word frequency related to these two topics increased an average of 95 percentage points due to inspection. These results are robust to visual examination of parallel trends, balance tests, and falsification tests using alternative outcomes.

These findings are relevant for U.S. education policy. Currently, school districts rely primarily on standardized tests for accountability purposes. Test-based accountability aims to improve standardized test results through incentives (Figlio & Loeb, 2011). In this context, schools are accountable for test results, but not their specific improvement actions. How schools choose to address low test scores is not emphasized. The sole focus on test results is associated with unintended

consequences, such as narrowed curriculum (e.g. Fitchett & Heafner, 2010; Jacob, 2005), gaming strategies to improve measured outcomes (Figlio & Loeb, 2011), and neglecting other critical aspects of learning that are not tested (Jacob, 2005; Rothstein et al., 2008). In contrast, school inspection not only creates incentives for improvement, but also provides specific feedback on school processes and outcomes (Ehren et al., 2013). This offers more nuanced information on school strengths and weakness to guide reforms. An increase in text related to influential inspection topics provides evidence that inspection was effective informing school reforms.

While this study does not address reform implementation, it is assumed that SIPs represent schools' intended reforms. A potential concern is that schools may present inauthentic goals in school plans, in order to please the district (Meyers & VanGronigen, 2019). This is unlikely to be the case due to several factors. First, test-based accountability remains the primary mechanism that principals are responsive to, as emphasized in interviews. Further, the SIPs emphasize standardized results since they are a state requirement relevant for Federal funding, which is linked to test performance. Second, the district does not track the reforms implemented after inspection; thus, there are not direct incentives to implement changes based on inspection feedback. Finally, the scope of inspection much broader than the two influential areas identified in these studies. These two areas emphasized by inspectors, nor by the district; thus, it does not seem to be the case that principals would be incentivized to mention these two areas in order to please the district.

These findings also contribute to the limited literature that has assessed the effect of inspection on school reform efforts (Cuckle et al., 1998; Dederling & Müller, 2011; Ehren & Visscher, 2008; J. Gray & Wilcox, 1995; Ouston et al., 1997; Verhaeghe et al., 2010). School planning is a crucial step in deciding the direction of reforms (Matthews & Sammons, 2004). Yet, prior literature has not evaluated the causal effect of inspection on planning. My study fills this gap, by taking advantage of the availability of SIPs before and after inspections take place and using

quasi-experimental methods. This led to finding the causal impact of school inspection on the focus of intended reforms on specific focus areas.

Overall, this study provides evidence regarding the potential of school inspection to guide school reforms. Heavy reliance on standardized test results offers limited insight into specific, beneficial reforms and create incentives to narrow the scope of reforms (e.g. Gagnon & Schneider, 2019). This study demonstrates the potential of on-site evaluation in the U.S. to inform the school planning process, providing a broad diagnosis of schools' strengths and weakness, identifying areas that hinder improvement, and involving school staff in the planning process.

APPENDICES

Appendix A – Codebook for Interviews to School Principals

1. Usefulness

1.1. Useful / New insights

1.1.1. Better prioritize

1.1.2. Reaffirm existing goals / Confirm diagnosis

1.1.3. Gain legitimacy

1.1.4. Somewhat useful

1.2. Not useful / Not relevant

5. Changes – What motivates them?

5.1. Principal initiatives / Staff initiatives

5.2. Inspections

5.2.1. Community Involvement

5.2.2. Climate & Culture conducive to Learning

5.2.3. Instructional Practices

5.2.4. Leadership

5.2.5. Professional Development

5.2.6. Support to Students

5.2.7. Teachers-Administrators Collaboration

5.2.8. Other Organizational Issues.

5.3. Test results, performance framework, evaluations

5.4. School supervisors

5.5. District (excluding 5.3. & 5.4)

5.6. Other sources

5.7. Did not implement changes based on inspection (explicit)

Appendix B – Content Analysis Dictionary

1) Instructional Practices Terms:

Active Learning, CCSS (Common Core State Standards) ,Check* for understanding, Class* inst*, Class size*, Class time, Common core, Conducive to learning, Consistent expectations, Coop*, Co-Teach*, Culturally responsive, Differentiat* instruction, Differentiat* learning, Differentiation, Direct instruction, Embedded assessment, Exit ticket*, Experiential learning, Feedback to students, Formative, Grouping, Growth mindset, Growth mind set, Individualize*, Individualizing, Inquiry based, Instructional method*, Instructional practices, Instructional strategies, Intentional, Lesson design, Lesson* plan*, Misconceptions, Mistakes, Misunderstanding, PBL (Project Based Learning), Pedagog*, Peer to peer, Plan* lesson*, Prior knowledge, Problem solving, Project based, Questioning, Quiz*, Real-Life, Reasoning, Regular*_Assess*, Re-Teaching, Rigor, Shelter*_Instruction, Small group*, Standard, Structured learning, Student centered, Targeted instruction, Teacher created, Teacher led, Thinking, Unite assesement*, Whole group, Time in class*

2) Culture & Climate conducive to Learning Terms:

Safe, Abuse, Academic culture, Addiction*, Alcohol, Attitude*, Behav*, Build relationships, Bully*, Class* climate, Class* culture, Classroom environment, Classroom environment*, Collaborative systems, Collective, Conflict*, Dean of culture, Drug*, Emotion*, Empath*, High expectations, Improvement culture, Instruction culture, Interaction*, Interpersonal, Learning culture, Learning environment*, Marijuana, Norms, PBIS (Positive Behavior Intervention Supports), Positive climate, Positive culture, Positive relationships, Positive school culture, Relationships between, Relationships with students, Respect, Respectful, Restorative, RJ (Restorative Justice), Routines, Rules, School's culture, School climate, School culture, School wide culture, SEAL (Social, Emotional, And

Academic Learning), SEL (Social And Emotional Learning), Student engagement, Student culture, Student voice, Suicide, Trauma*, Truan*, Trust, Wellness, Whole child.

Appendix C –Most Frequent Phrases on School Improvement Plans – School Years 2016-17 and 2018-19

ACADEMIC_PERFORMANCE STUDENT_ENGAGEMENT FAMILY_ENGAGEMENT LITERACY AND MATH
CONCURRENT_ENROLLMENT SCHOOL_COMMUNITY CONTENT AREAS
FREE_AND_REDUCED_LUNCH ACADEMIC_GROWTH DATA_DRIVEN
AFRICAN_AMERICAN STUDENTS_OF_COLOR LEARNING_ENVIRONMENT
SCHOOL_PERFORMANCE_FRAMEWORK CULTURE_AND_CLIMATE
COMMON_CORE LANGUAGE_ARTS
AFTER_SCHOOL POST_SECONDARY
IMPROVEMENT_PLAN SCHOOL_WIDE DATA_ANALYSIS
EARLY_LITERACY
ENGLISH_LANGUAGE_LEARNERS
PROFESSIONAL_DEVELOPMENT
ASSESSMENT_DATA **SPECIAL_EDUCATION**
COMMUNITY_ENGAGEMENT
EARLY_COLLEGE SCHOOL_CULTURE
NATIVE_LANGUAGE IMPROVEMENT_STRATEGIES
STAFF_MEMBERS STUDENT_ACHIEVEMENT
SOCIAL_EMOTIONAL READ_ACT SOCIAL_STUDIES
REDUCED_LUNCH INSTRUCTIONAL_LEADERSHIP_TEAM
INTERIM_ASSESSMENTS COLLABORATIVE_SCHOOL_COMMITTEE
PARENT_INVOLVEMENT TEACHER LEADERS COMMUNITY_MEMBERS
HIGH_EXPECTATIONS INSTRUCTIONAL_MODELS PROFESSIONAL_LEARNING

Appendix D – Test Score Results

	School Year 2016-17		School Year 2018-19	
	Not Inspected	Inspected	Not Inspected	Inspected
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Test results				
English Language Arts	737.9*** (20.85)	725.6*** (15.39)	741.4*** (20.11)	730.3*** (13.69)
Math	732.9*** (18.18)	722.9*** (12.14)	734.8*** (18.47)	722.3*** (11.85)
n	102	26	102	26

***: significant difference at the 1% level from two-sample t tests between inspected and not inspected schools.

Appendix E – Dictionary for Placebo Test

Other Inspection Related Terms:

African American, After school, American Indian, At risk, Authorizer, Autism, Bilingual, Bilingual parent advisory committee, Biliteracy, Black student*, Candidate*, Chinese, Club*, Coach*, Collaborative planning, Community event*, Community partnership, Compliance, Conference*, Decision-Mak*, Department meeting*, Disab*, Distributed leadership, Dual language, Educator* need*, Effective teacher*, ELD, ELL, English as a second language, English language development, English language learner*, ESL, Extra-curricular, Faculty input, Families, Family, Father*, Financial, Food Service*, Frequent communication*, Grade level meeting*, Granparent*, Guardian*, High quality teach*, Hire*, Hispanic, Home visit*, Immigrant*, In need, Intervention*, Job embedded, Language acquisition, Language immersion, Latin*, Lead teacher*, Leadership meeting*, Leadership model, Leadership support, Lesson observation*, Lesson planning, Meet frequently, Mentor*, Minorit*, Mother*, Multilingual, Native*, Neighbor*, Observ* other teacher*, Observ* teacher*, Open communication*, Operational, Organizational goal*, Parent*, Professional development, Professional growth, Professional learning, Professional standard*, Race*, Recruit*, Reflective process*, Refugee*, Response to Intervention , Retain*, Reten*, School event*, School leadership team*, School staff, School* operation*, SLT*, SOC, Spanish speaking, Special education, Special needs, Sports, Staff evaluation*, Staff input, Staff meeting*, Staff review*, Staff superv*, Staff support, Strategic conversation*, Strategic plan*, Student* need*, Student* of color, Student* support, Summer program*, Supplemental services, Support to student*, Support* staff, System*, Teacher meeting*, Teaching staff, Team meeting*, TNLI, Training*, Transitional native, language instruction, Turnover, Tutor*, Underrepresented, Volunteer*, White student*, Workshop*, Youth center*

REFERENCES

REFERENCES

- Ahuvia, A. (2001). Traditional, Interpretive, and Reception Based Content Analyses: Improving the Ability of Content Analysis to Address Issues of Pragmatic and Theoretical Concern. *Social Indicators Research*, 54, 139–172. [https://doi.org/10.1108/0167-4544\(200105\)54:1<139::AID-SIR139>3.0.CO;2](https://doi.org/10.1108/0167-4544(200105)54:1<139::AID-SIR139>3.0.CO;2)
- Allen, R., & Burgess, S. (2012). *How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England* (No. 12; 87).
- Altrichter, H., & Kemethofer, D. (2015). Does accountability pressure through school inspections promote school improvement? *School Effectiveness and School Improvement*, 26(1), 32–56. <https://doi.org/10.1080/09243453.2014.927369>
- Apple, M. (2005). Education, markets, and an audit culture. *Critical Quarterly*, 47(1–2), 11–29. <https://doi.org/doi:10.1111/j.0011-1562.2005.00611>
- Armenakis, A., Bernerth, J., Pitts, J., & Walker, H. (2007). Organizational Change Recipients' Beliefs Scale. *The Journal of Applied Behavioral Science*, 43(4), 481–505. <https://doi.org/10.1177/0021886307303654>
- Armenakis, A., & Harris, S. (2009). Reflections: our Journey in Organizational Change Research and Practice. *Journal of Change Management*, 9(2), 127–142. <https://doi.org/10.1080/14697010902879079>
- Armenakis, A., Harris, S., Cole, M., Fillmer, L., & Self, D. (2007). A Top Management Team's Reactions to Organizational Transformation: The Diagnostic Benefits of Five Key Change Sentiments. *Journal of Change Management*, 7(3–4), 273–290. <https://doi.org/10.1080/14697010701771014>
- Armstrong, J. (1982). The value of formal planning for strategic decisions: review of empirical research. *Strategic Management Journal*, 3, 197–211.
- Ball, S., & Bowe, R. (1992). Subject departments and the 'implementation' of National Curriculum policy: an overview of the issues. *Journal of Curriculum Studies*, 24(2), 97–115. <https://doi.org/10.1080/0022027920240201>
- Barber, M. (2005). The virtue of accountability: System redesign, inspection, and incentives in the era of informed professionalism. *Journal of Education*, 185(1), 7–38. <https://doi.org/10.1177/002205740518500102>
- Baxter, J. A. (2013). Professional inspector or inspecting professional? Teachers as inspectors in a new regulatory regime for education in England. *Cambridge Journal of Education*, 43(4), 467–485. <https://doi.org/10.1080/0305764X.2013.819069>
- Behnke, K., & Steins, G. (2017). Principals' reactions to feedback received by school inspection: A longitudinal study. *Journal of Educational Change*, 18(1), 77–106.

<https://doi.org/10.1007/s10833-016-9275-7>

- Bengston, D., & Xu, Z. (1995). *Changing National Forest Values: a content analysis - Research Paper NC-323*. http://www.nrs.fs.fed.us/pubs/rp/rp_nc323.pdf
- Berry, F. S., & Wechsler, B. (1995). State agencies' experience with strategic planning: findings from a national survey. *Public Administration Review*, 55(2), 159. <https://doi.org/10.2307/977181>
- Bitan, K., Haep, A., & Steins, G. (2014). School inspections still in dispute – an exploratory study of school principals' perceptions of school inspections. *International Journal of Leadership in Education*, 18(4), 1–22. <https://doi.org/10.1080/13603124.2014.958199>
- Bloem, S. (2015). The OECD Directorate for Education as an independent knowledge producer through Pisa. In H. G. Kotthoff & E. Klerides (Eds.), *Governing Educational Spaces* (pp. 169–185). SensePublishers. https://doi.org/10.1007/978-94-6300-265-3_10
- Brier, A., & Hopp, B. (2011). Computer assisted text analysis in the social sciences. *Quality & Quantity*, 45(1), 103–128. <https://doi.org/10.1007/s11135-010-9350-8>
- Chabbott, C., & Elliott, E. J. (2003). Understanding others, educating ourselves: Getting more from international comparative studies in education. In *Social Sciences*. <https://doi.org/10.17226/10622>
- Chun, Y. H., & Rainey, H. G. (2005). Goal ambiguity and organizational performance in U.S. federal agencies. *Journal of Public Administration Research and Theory*, 15(4), 529–557. <https://doi.org/10.1093/jopart/mui030>
- Clarke, J., & Ozga, J. (2011). Governing by inspection? Comparing school inspection in Scotland and England. *Social Policy Association Conference*, 25.
- Coburn, C. (2001). Beyond decoupling: Rethinking the relationship between the institutional environment and the classroom. *Sociology of Education*, 77, 211–244. <https://doi.org/10.1177/003804070407700302>
- Coburn, C. (2005). Shaping teacher sensemaking: School leaders and the enactment of reading policy. *Educational Policy*, 19(3), 476–509. <https://doi.org/10.1177/0895904805276143>
- Cole, M. S., Harris, S., & Bernerth, J. B. (2006). Exploring the implications of vision, appropriateness, and execution of organizational change. *Leadership & Organization Development Journal*, 27(5), 352–367. <https://doi.org/10.1108/01437730610677963>
- Concurso de Supervisores Rio Negro, (2013).
- Resolución del Consejo Provincial de Educación de Río Negro N 1053, Pub. L. No. 1053 (1994).
- Conway, M. (2006). The Subjective Precision of Computers: A Methodological Comparison with Human Coding in Content Analysis. *Journalism & Mass Communication Quarterly*,

- 83(1), 186–200. <https://doi.org/10.1177/107769900608300112>
- Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Cuckle, P., Hodgson, J., & Broadhead, P. (1998). Investigating the relationship between OFSTED Inspections and school development planning. *School Leadership & Management, 18*(2), 271–283. <https://doi.org/10.1080/13632439869691>
- Darling-Hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Podolsky, A., & Stosich, E. L. (2016). *Pathways to new accountability through the Every Student Succeeds Act*. <http://learningpolicyinstitute.org/our-work/publications-resources/%0Apathways-new-accountability-every-student-succeeds-act>.
- De Vries, H., Elliott, M. N., Kanouse, D. E., & Teleki, S. S. (2008). Using Pooled Kappa to Summarize Interrater Agreement across Many Items. *Field Methods, 20*(3), 272–282. <https://doi.org/10.1177/1525822X08317166>
- de Wolf, I., & Janssens, F. (2007). Effects and side effects of inspections and accountability in education: An overview of empirical studies. *Oxford Review of Education, 33*(3), 379–396. <https://doi.org/10.1080/03054980701366207>
- Dedering, K., & Müller, S. (2011). School improvement through inspections? First empirical insights from Germany. *Journal of Educational Change, 12*(3), 301–322. <https://doi.org/10.1007/s10833-010-9151-9>
- Dedering, K., & Sowada, M. G. (2017). Reaching a conclusion—procedures and processes of judgement formation in school inspection teams. *Educational Assessment, Evaluation and Accountability, 29*(1), 5–22. <https://doi.org/10.1007/s11092-016-9246-9>
- Deng, Q., Hine, M., Ji, S., & Sur, S. (2019). Inside the black box of dictionary building for text analytics: a design science approach. *Journal of International Technology and Information Management, 27*(3), 119–159.
- Doud, J. (1995). Planning for school improvement: A curriculum model for school based evaluation. *Peabody Journal of Education, 70*, 175–187.
- Edgerton, A. K. (2019). The essence of ESSA: More control at the district level? *Phi Delta Kappan, 101*(2), 14–17. <https://doi.org/10.1177/0031721719879148>
- Education Inspectorate - Ministry of Education, C. and S. (2010). *Risk-based Inspection as of 2009 - Primary and Secondary Education*.
- Education Inspectorate - Ministry of Education, C. and S. (2017a). *Inspection framework primary education*.
- Education Inspectorate - Ministry of Education, C. and S. (2017b). *Inspection framework secondary education*.

- Ehren, M. (2016a). *Methods and modalities of effective school inspections* (M. Ehren (ed.)). Springer International Publishing. <https://doi.org/10.1007/978-3-319-31003-9>
- Ehren, M. (2016b). *Methods and Modalities of Effective School Inspections*. In M. C.M. Ehren (Ed.), *Methods and Modalities of Effective School Inspections*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-31003-9>
- Ehren, M., Altrichter, H., McNamara, G., & O'Hara, J. (2013). Impact of school inspections on improvement of schools—describing assumptions on causal mechanisms in six European countries. *Educ Asse Eval Acc*, 25, 3–43. <https://doi.org/10.1007/s11092-012-9156-4>
- Ehren, M., Gustafsson, J.-E., Altrichter, H., Skedsmo, G., Kemethofer, D., & Huber, S. (2015). Comparing effects and side effects of different school inspection systems across Europe. *Comparative Education*, 51(3), 375–400. <https://doi.org/10.1080/03050068.2015.1045769>
- Ehren, M., Perryman, J., & Shackleton, N. (2015a). School Effectiveness and School Improvement. *School Effectiveness and School Improvement - An International Journal of Research, Policy and Practice*, 26(2), 296–327.
- Ehren, M., Perryman, J., & Shackleton, N. (2015b). Setting expectations for good education: how Dutch school inspections drive improvement. *School Effectiveness and School Improvement*, 26(2), 296–327. <https://doi.org/10.1080/09243453.2014.936472>
- Ehren, M., & Shackleton, N. (2016). Risk-based school inspections: impact of targeted inspection approaches on Dutch secondary schools. *Educational Assessment, Evaluation and Accountability*, 28(4), 299–321. <https://doi.org/10.1007/s11092-016-9242-0>
- Ehren, M., & Visscher, A. (2006). TOWARDS A THEORY ON THE IMPACT OF SCHOOL INSPECTIONS. *British Journal of Educational Studies*, 54(1), 51–72. <https://doi.org/10.1111/j.1467-8527.2006.00333.x>
- Ehren, M., & Visscher, A. (2008). THE RELATIONSHIPS BETWEEN SCHOOL INSPECTIONS, SCHOOL CHARACTERISTICS AND SCHOOL IMPROVEMENT. *British Journal of Educational Studies*, 56(2), 205–227. <https://doi.org/10.1111/j.1467-8527.2008.00400.x>
- Fernandez, K. E. (2011). Evaluating school improvement plans and their affect on academic performance. *Educational Policy*, 25(2), 338–367. <https://doi.org/10.1177/0895904809351693>
- Figlio, D., & Loeb, S. (2011). School accountability. In *Handbook of the Economics of Education* (pp. 383–421).
- Fitchett, P., & Heafner, T. (2010). A national perspective on the effects of high-stakes testing and standardization on elementary social studies marginalization. *Theory & Research in Social Education*, 38(1), 114–130. <https://doi.org/10.1080/00933104.2010.10473418>
- Gagnon, D. J., & Schneider, J. (2019). Holistic school quality measurement and the future of

- accountability: Pilot-test results. *Educational Policy*, 33(5), 734–760.
<https://doi.org/10.1177/0895904817736631>
- Gilroy, P., & Wilcox, B. (1997). OFSTED, criteria and the nature of social understanding: A Wittgensteinian critique of the practice of educational judgement. *British Journal of Educational Studies*, 45(1), 22–38. <https://doi.org/10.1111/1467-8527.00034>
- Gioia, D., Thomas, J., Clark, S., & Chittipeddi, K. (1994). Symbolism and strategic change in academia: The dynamics of sensemaking and influence. *Organization Science*, 5(3), 363–383. <https://doi.org/10.1287/orsc.5.3.363>
- Glazerman, S. (2016). *The false dichotomy of school inspection*. Mathematica Policy Research - Blog Post. <https://www.mathematica-mpr.com/commentary/the-false-dichotomy-of-school-inspections>
- Gray, C., & Gardner, J. (1999). The impact of school inspections. *Oxford Review of Education*, 25(4), 455–468. <https://doi.org/10.1080/030549899103928>
- Gray, J., & Wilcox, B. (1995). In the aftermath of inspection: the nature and fate of inspection report recommendations. *Research Papers in Education*, 10(1), 1–18.
<https://doi.org/10.1080/0267152950100102>
- Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11(13), 255–274.
- Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267–297.
<https://doi.org/10.1093/pan/mps028>
- Grimolizzi-Jensen, C. J. (2018). Organizational change: Effect of motivational interviewing on readiness to change. *Journal of Change Management*, 18(1), 54–69.
<https://doi.org/10.1080/14697017.2017.1349162>
- Gustafsson, J.-E., Ehren, M., Conyngham, G., McNamara, G., Altrichter, H., & O'Hara, J. (2015). From inspection to quality: Ways in which school inspection influences change in schools. *Studies in Educational Evaluation*, 47, 47–57.
<https://doi.org/10.1016/j.stueduc.2015.07.002>
- Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. *Educational Administration, Policy, and Reform: Research and Measurement*, 3, 153–188.
- Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2), 297–327.
<https://doi.org/10.1002/pam.20091>
- Herscovitch, L., & Meyer, J. P. (2002). Commitment to organizational change: Extension of a

- three-component model. *Journal of Applied Psychology*, 87(3), 474–487.
<https://doi.org/10.1037/0021-9010.87.3.474>
- Hill, H. (2001). Policy is not enough: language and the interpretation of State standards. *American Educational Research Journal*, 38(2), 289–318.
<https://doi.org/10.3102/00028312038002289>
- Hines, R. T. (2017). *An Exploration of the Effects of School Improvement Planning and Feedback Systems: School Performance in North Carolina*.
- Holt, D., Armenakis, A., Feild, H., & Harris, S. (2007). Readiness for Organizational Change. *The Journal of Applied Behavioral Science*, 43(2), 232–255.
<https://doi.org/10.1177/0021886306295295>
- Husfeldt, V. (2011). Wirkungen und wirksamkeit der externen schulevaluation; uberblick zum stand der forschung [The impact of school inspection - Does it really work? State of research]. *Zeitschrift Für Erziehungswissenschaft*, 14(2), 259–282.
<https://doi.org/10.1007/s11618-011-0204-5>
- Hussain, I. (2015). Subjective Performance Evaluation in the Public Sector Evidence from School Inspections. *The Journal of Human Resources*, 50(1), 189–221.
- Jacob, B. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5–6), 761–796.
<https://doi.org/10.1016/j.jpubeco.2004.08.004>
- Jones, K., & Tymms, P. (2014). Ofsted’s role in promoting school improvement: the mechanisms of the school inspection system in England. *Oxford Review of Education*, 40(3), 315–330.
- Jones, K., Tymms, P., Kemethofer, D., O’Hara, J., McNamara, G., Huber, S., Myrberg, E., Skedsmo, G., & Greger, D. (2017). The unintended consequences of school inspection: the prevalence of inspection side-effects in Austria, the Czech Republic, England, Ireland, the Netherlands, Sweden, and Switzerland. *Oxford Review of Education*, 43(6), 805–822.
<https://doi.org/10.1080/03054985.2017.1352499>
- Kaplan, S., & Orlikowski, W. J. (2013). Temporal Work in Strategy Making. *Organization Science*, 24(4), 965–995. <https://doi.org/10.1287/orsc.1120.0792>
- Klein, A. (2016). *School inspections offer a diagnostic look at quality*. Education Week.
<https://www.edweek.org/ew/articles/2016/09/28/school-inspections-offer-a-diagnostic-look-at.html>
- Klerks, M. (2012). *The effect of school inspections: a systematic review*. <http://janbri.nl/wp-content/uploads/2014/12/ORD-paper-2012-Review-Effect-School-Inspections-MKLERKS.pdf>
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A

- historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>
- Koretz, D. (2008). *Measuring up*. Harvard University Press.
- Krippendorff, K. (2013). *Content analysis: An introduction to Its methodology* (3rd ed.). SAGE Publications.
- Ladd, H. F. (2016). *Now is the time to experiment with inspections for school accountability*. Brookings. <https://www.brookings.edu/blog/brown-center-chalkboard/2016/05/26/now-is-the-time-to-experiment-with-inspections-for-school-accountability/>
- Ladd, H. F. (2017). NCLB: RESPONSE TO JACOB. *Journal of Policy Analysis and Management*, 36(2), 477–480. <https://doi.org/10.1002/pam.21979>
- Ladd, H. F., & Figlio, D. (2008). School accountability and student achievement. In *Handbook of research in education finance and policy* (pp. 166–182).
- Lee, J., & Fitz, J. (1997). HMI and OFSTED: evolution or revolution in school inspection. *British Journal of Educational Studies*, 45(1), 39–52. <https://doi.org/10.1111/1467-8527.00035>
- Lewin, A. Y., & Minton, J. W. (1986). Determining Organizational Effectiveness: Another Look, and an Agenda for Research. *Management Science*, 32(5), 514–538. <https://doi.org/10.1287/mnsc.32.5.514>
- Lindgren, J. (2015). The front and back stages of swedish school inspection: opening the black box of judgment. *Scandinavian Journal of Educational Research*, 59(1), 58–76. <https://doi.org/10.1080/00313831.2013.838803>
- Luginbuhl, R., Webbink, D., & de Wolf, I. (2009). Do Inspections Improve Primary School Performance? *Analysis*, 31(3), 221–237. <https://doi.org/10.3102/0162373709338315>
- Maitlis, S. (2005). The Social Processes of Organizational Sensemaking. *The Academy of Management Journal*, 48(1), 21–49. <https://doi.org/10.2307/20159639>
- Maitlis, S., & Christianson, M. (2014). Sensemaking in organizations: Taking stocks and moving forward. *The Academy of Management Annals*, 8(1), 57–125. <https://doi.org/10.1080/19416520.2014.873177>
- March, J. G., & Olsen, J. P. (2011). The Logic of Appropriateness. In R. E. Goodin (Ed.), *The Oxford Handbook of Political Science* (pp. 1–22). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199604456.013.0024>
- Mathis, W., & Trujillo, T. (2016). *Lessons from NCLB for the Every Student Succeeds Act*. <http://nepc.colorado.edu/%0Apublication/lessons-from-NCLB>
- Matthews, P., & Sammons, P. (2004). *Improvement through inspection. An evaluation of the*

impact of Ofsted's work. Ofsted.

- Matthews, Peter, Holmes, J. R., Vickers, P., & Corporaal, B. (1998). Aspects of the reliability and validity of school inspection judgements of teaching quality. *Educational Research and Evaluation, 4*(2), 167–188. <https://doi.org/10.1076/edre.4.2.167.6959>
- McDonnell, L. (2008). The politics of educational accountability: Can the clock be turned back? In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability*. Routledge.
- McDonnell, L. (2013). Educational accountability and policy feedback. *Educational Policy, 27*(2), 170–189. <https://doi.org/10.1177/0895904812465119>
- Meyers, C. V., & VanGronigen, B. A. (2019). A lack of authentic school improvement plan development. *Journal of Educational Administration, 57*(3), 261–278. <https://doi.org/10.1108/JEA-09-2018-0154>
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3th ed.). SAGE Publications.
- Millett, A., & Johnson, D. C. (1998). Expertise or “baggage”? What helps inspectors to inspect primary mathematics? *British Educational Research Journal, 24*(5), 503–518. <https://doi.org/10.1080/0141192980240502>
- Mintrop, H., MacLellan, A. M., & Quintero, M. F. (2001). School improvement plans in schools on probation: A comparative content analysis across three accountability systems. *Educational Administration Quarterly, 37*(2), 197–218. <https://doi.org/10.1177/00131610121969299>
- Morse, J. (2010). Procedures and practice of mixed method design - Maintaining control, rigor, and complexity. In A. M. Tashakkori & C. B. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 339–352). SAGE Publications.
- Neuendorf, K. A. (2017). *The content analysis guidebook*. SAGE Publications, Inc. <https://doi.org/10.4135/9781071802878>
- Nusche, D., Braun, H., Halász, G., & Santiago, P. (2014). *OECD Reviews of Evaluation and Assessment in Education: Netherlands 2014*. OECD. <https://doi.org/10.1787/9789264211940-en>
- OECD, O. for E. C. and D. (2015). *Education at a glance 2015 - OECD Indicators*. <https://doi.org/10.1787/19991487>
- Ouston, J., Fidler, B., & Earley, P. (1997). What do schools do after OFSTED school inspections-or before? *School Leadership & Management, 17*(1), 95–104. <https://doi.org/10.1080/13632439770195>
- Penninckx, M., & Vanhoof, J. (2015). Insights gained by schools and emotional consequences of

- school inspections. A review of evidence. *School Leadership & Management*, 35(5), 477–501. <https://doi.org/10.1080/13632434.2015.1107036>
- Penninckx, M., Vanhoof, J., De Maeyer, S., & Van Petegem, P. (2014). Exploring and explaining the effects of being inspected. *Educational Studies*, 40(4), 456–472. <https://doi.org/10.1080/03055698.2014.930343>
- Penninckx, M., Vanhoof, J., De Maeyer, S., & Van Petegem, P. (2015). Effects and side effects of Flemish school inspection. *Educational Management Administration & Leadership*. <https://doi.org/10.1177/1741143215570305>
- Perryman, J. (2007). Inspection and emotion. *Cambridge Journal of Education*, 37(2), 173–190. <https://doi.org/10.1080/03057640701372418>
- Perryman, J. (2009). Inspection and the fabrication of professional and performative processes. *Journal of Education Policy*, 24(5), 611–631.
- Phillips, D., & Schweisfurth, M. (2014). *Comparative and international education - An introduction to theory, methods, and practice* (2nd ed.). Group, Continuum International Publishing.
- Piderit, S. K. (2000). Rethinking resistance and recognizing ambivalence: A multidimensional view of attitudes toward an organizational change. *The Academy of Management Review*, 25(4), 783. <https://doi.org/10.2307/259206>
- Pond, S., Armenakis, A., & Green, S. (1984). The Importance of Employee Expectations in Organizational Diagnosis. *The Journal of Applied Behavioral Science*, 20(2), 167–180. <https://doi.org/10.1177/002188638402000207>
- Porac, J. F., Thomas, H., & Baden-Fuller, C. (1989). COMPETITIVE GROUPS AS COGNITIVE COMMUNITIES: THE CASE OF SCOTTISH KNITWEAR MANUFACTURERS. *Journal of Management Studies*, 26(4), 397–416. <https://doi.org/10.1111/j.1467-6486.1989.tb00736.x>
- Portz, J., & Beauchamp, N. (2020). Educational Accountability and State ESSA Plans. *Educational Policy*, 089590482091736. <https://doi.org/10.1177/0895904820917364>
- Ravitch, D. (2016). *The death and life of the great American school system: How testing and choice are undermining education*. Basic Books.
- Redding, C., & Searby, L. (2020). The Map Is Not the Territory: Considering the Role of School Improvement Plans in Turnaround Schools. *Journal of Cases in Educational Leadership*, 23(3), 63–75. <https://doi.org/10.1177/1555458920938854>
- Riffe, D., Lacy, S., & Fico, F. (2014). *Analyzing media messages: Using quantitative content analysis in research*. Routledge.
- Rigby, J. G. (2015). Principals' sensemaking and enactment of teacher evaluation. *Journal of*

- Educational Administration*, 53(3), 374–392. <https://doi.org/10.1108/JEA-04-2014-0051>
- Rosenthal, L. (2004). Do school inspections improve school quality? Ofsted inspections and school examination results in the UK. *Economics of Education Review*, 23, 143–151.
- Rothstein, R., Jacobsen, R., & Wilder, T. (2008). *Grading education: Getting accountability right*. Economic Policy Institute and Teacher College Press.
- Rouleau, L. (2005). Micro-practices of strategic sensemaking and sensegiving: How middle managers interpret and sell change every day. *Journal of Management Studies*, 42(7), 1413–1441.
- Rutz, S., Mathew, D., Robben, P., & Bont, A. (2017). Enhancing responsiveness and consistency: Comparing the collective use of discretion and discretionary room at inspectorates in England and the Netherlands. *Regulation & Governance*, 11(1), 81–94. <https://doi.org/10.1111/regg.12101>
- Ryan, K., Gandha, T., & Ahn, J. (2013). School self-evaluation and inspection for improving U.S. schools? In *National Education Policy Center*. <http://nepc.colorado.edu/publication/school-self-evaluation>
- Sandberg, J., & Tsoukas, H. (2015). Making sense of the sensemaking perspective: Its constituents, limitations, and opportunities for further development. *Journal of Organizational Behavior*, 36(S1), S6–S32. <https://doi.org/10.1002/job.1937>
- Scheerens, J., Ehren, M., Slegers, P., & Leeuw, R. de. (2012). *OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes*.
- Shaw, I., Newton, D. P., Aitkin, M., & Darnell, R. (2003). Do OFSTED Inspections of Secondary Schools Make a Difference to GCSE Results? *British Educational Research Journal*, 29(1), 63–75.
- Spillane, J. P. (1999). External reform initiatives and teachers' efforts to reconstruct their practice: The mediating role of teachers' zones of enactment. *Journal of Curriculum Studies*, 31(2), 1–33. <https://doi.org/10.1080/002202799183205>
- Spillane, J. P., Parise, L. M., & Sherer, J. Z. (2011). Organizational Routines as Coupling Mechanisms. *American Educational Research Journal*, 48(3), 586–619. <https://doi.org/10.3102/0002831210385102>
- Spillane, J. P., Reiser, 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 education policy implementation* (pp. 47–64). State University of New York Press, Albany.
- Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of Educational Research*, 72(3), 387–341. <https://doi.org/10.3102/00346543072003387>

- Stiglitz, J. (2000). *Economics of the public sector* (3rd ed.). Norton.
- Strunk, K. O., Marsh, J. A., Bush-Mecenas, S., & Duque, M. R. (2016). The Best Laid Plans. *Educational Administration Quarterly*, *52*(2), 259–309. <https://doi.org/10.1177/0013161X15616864>
- Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating qualitative and quantitative approaches in the social and behavioral sciences*. SAGE.
- Teddlie, C., & Yu, F. (2007). Mixed methods sampling : A typology with examples. *Journal of Mixed Methods Research*, *1*(1), 77–100. <https://doi.org/10.1177/1558689806292430>
- UNESCO. (2017). *Global Education Monitoring Report - Accountability in education: Meeting our commitments*.
- van Bruggen, J. C. (2010). *Inspectorates of Education in Europe; some comparative remarks about their tasks and work*.
- van der Sluis, M. E., Reezigt, G. J., & Borghans, L. (2017). Implementing New Public Management in Educational Policy. *Educational Policy*, *31*(3), 303–329.
- Vavrus, F. K., & Bartlett, L. (2016). *Rethinking case study research: A comparative approach* (1st ed.). Routledge.
- Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2010). Using school performance feedback: perceptions of primary school principals. *School Effectiveness and School Improvement*, *21*(2), 167–188. <https://doi.org/10.1080/09243450903396005>
- Visscher, A. J., & Coe, R. (2003). School performance feedback Systems: conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement*, *14*(3), 321–349. <https://doi.org/10.1076/sesi.14.3.321.15842>
- Weick, K. E. (1995). *Sensemaking in organizations*. SAGE Publications.
- Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, *16*(4), 409–421. <https://doi.org/10.1287/orsc.1050.0133>
- Weiner, B. J. (2009). A theory of organizational readiness for change. *Implementation Science*, *4*(1), 67. <https://doi.org/10.1186/1748-5908-4-67>
- Woods, P., & Jeffrey, B. (1998). Choosing Positions: Living the Contradictions of OFSTED. *British Journal of Sociology of Education*, *19*(4), 547–570. <https://doi.org/10.1080/0142569980190406>