Educational Evaluation and Policy Analysis
December 2019, Vol. 41, No. 4, pp. 510–536
DOI: 10.3102/0162373719869318
Article reuse guidelines: sagepub.com/journals-permissions
© 2019 AERA. http://eepa.aera.net

Using a Text-as-Data Approach to Understand Reform Processes: A Deep Exploration of School Improvement Strategies

Min Sun
University of Washington
Jing Liu
Brown University
Junmeng Zhu
Zachary LeClair
University of Washington

Although program evaluations using rigorous quasi-experimental or experimental designs can inform decisions about whether to continue or terminate a given program, they often have limited ability to reveal the mechanisms by which complex interventions achieve their effects. To illuminate these mechanisms, this article analyzes novel text data from thousands of school improvement planning and implementation reports from Washington State, deploying computer-assisted techniques to extract measures of school improvement processes. Our analysis identified 15 coherent reform strategies that varied greatly across schools and over time. The prevalence of identified reform strategies was largely consistent with school leaders' own perceptions of reform priorities via interviews. Several reform strategy measures were significantly associated with reductions in student chronic absenteeism and improvements in student achievement. We finally discuss the opportunities and pitfalls of using novel text data to study reform processes.

Keywords: *text as data, school improvement, reform processes*

In the past two decades, the development of experimental and quasi-experimental research designs in education research has significantly improved researchers' abilities to attribute observed changes in outcomes to specific policies or programs. Although such research can inform policymakers' decisions about whether to expand or terminate a certain program, it often has little influence on the theories of change employed by practitioners to support successful program implementation in schools and districts (Singer, 2018), mainly because this type of research is limited in its ability to reveal the mechanisms by which complex interventions achieve their effects (Hedges, 2018). To make education research more useful to practitioners, researchers need to go beyond determining whether a program works and uncover what processes make the program work and how (Hedges, 2018; Singer, 2018). To further this endeavor, the current article explores an emerging method for analyzing a relatively untapped source of textual data on school reform activities.

Prior studies in education evaluation and policy analysis have used various approaches to investigate the contexts and mechanisms of change, but each of these approaches has its own limitations. For example, education researchers often use administrative data on teacher characteristics and student demographics to study variation in program effects across school contexts or student subgroups. However, such studies depend on the availability of these measures in administrative data sets and are unable to fully probe into the actual strategies and processes of change. Some researchers have recently advanced the use of mediation analysis to study change pathways (e.g., Hong & Nomi, 2012; Raudenbush, Reardon, & Nomi, 2012; Reardon & Raudenbush, 2013; Weiss, Bloom, & Brock, 2014), but the assumptions and data requirements to conduct mediation analysis in a well-designed multisite experiment are not always easy to establish. Another approach to examining change mechanisms is through fieldwork (e.g., interviews and observations), but this type of research is expensive to conduct and it is often difficult to quantify qualitative data collected at a large scale.

To address some of these limitations, in the current article, we propose an alternative approach to examining change mechanisms using a new form of program implementation artifacts: texts and documents. In school improvement efforts, whether required by the district or state or voluntarily undertaken by individual schools, schools often use written reports to establish visions, design reform strategies, coordinate efforts among key stakeholders, and monitor reform implementation (Strunk, Marsh, Bush-Mecenas, & Duque, 2016). These reports contain rich information on the planning and implementation of school improvement efforts and often include valuable explanations of how and why certain programs work. Yet, these reports are rarely analyzed quantitatively and systematically because the conventional approach to document analysis—using human annotators to code the unstructured text in these reports—is often time consuming and costly (Strunk et al., 2016).

Recent developments in computer-assisted text analysis offer promising solutions to such challenging issues. Originally developed in computer science, these methods have more recently been adopted by social scientists, particularly political scientists, to significantly advance theory development. Just to name a few examples, Wilkerson, Smith, and Stramp (2015) investigated "text reuse" methods as a means for tracing the progress of policy ideas in legislation, providing new insights into the lawmaking process. Kim (2017) investigated the contents of trade bills using latent Dirichlet allocation (LDA); his findings challenged the common focus on industry-level lobbying preferences. Grimmer, Messing, and Westwood (2012) used super-vised classification methods to analyze more than 170,000 House press releases and examine legislators' credit-claiming behavior, wherein legislators associate themselves with spending in their constituent districts to cultivate votes. Such computer-assisted techniques (e.g., text reuse, topic modeling, classification methods) allow for systematic analysis of large-scale text collections without massive funding support (Grimmer & Stewart, 2013). However, the application of such methods is still sparse in education policy research.

In this article, we apply text analyses, particularly LDA, to identify key, fine-grained measures of school improvement strategies and schools' differential priorities at a large scale during the era of No Child Left Behind waivers and federal School Improvement Grants (SIGs) in Washington State. After comparing several model specifications, we identified 20 reform strategies that emerged from the data, which varied greatly across schools by reform type and over time. Our expert human coders verified each identified reform strategy and concluded that 15 of these 20 measures were conceptually coherent. Using interview data, we also found that the identified reform strategies were largely consistent with school leaders' own perceptions of reform priorities. Finally, we illustrated the predictive relations of these reform strategy measures by showing that several measures were significantly associated with the reductions in student chronic absenteeism and the improvements in student achievement. Together, this descriptive study demonstrates the potential of using text-as-data approaches to study education policy processes, and identifies a few school reform strategies that are significantly associated with the improvement in student outcomes.

In the next section, we review the emerging body of education research using text analysis and discuss the limitations of these studies. We then describe the policy and implementation background of school turnaround efforts in Washington State, along with our sample, measures, and text-as-data methods. Finally, we summarize the findings and discuss the potential benefits and drawbacks of using this new form of data in education evaluation and policy analysis.

Text Analysis in Education Research

Text-as-data methods are a promising tool for education policy research, especially for systematically quantifying conventionally hard-to-measure, yet important, schooling processes and individual attributes. This section discusses two new applications of text-as-data methods in education research, with the understanding that these applications are limited in both research areas and methodological rigor.

First, researchers have begun to use text-as-data methods to measure latent dispositions, attitudes, and beliefs of students and teachers. For example, Beattie, Laliberté, and Oreopoulos (2018) used a topic model to analyze college students' responses to open-ended questions, such as what kind of person they aspire to be in their life. The topics derived from the analysis were used as proxies of students' expectations and aspirations. The authors found significant differences between high-performing students and their low-performing peers on these nonacademic measures. In a similar vein, Penner, Rochmes, Liu, Solanki, and Loeb (2019) used a structural topic model to code teachers' values and beliefs about student achievement gaps by using essays written by more than 10,000 job applicants at an urban California school district. They found that certain themes were systematically correlated with applicants' characteristics, the schools they were applying for, and their hiring outcomes.

Second, some researchers have applied text-as-data methods to investigate microclassroom processes, including peer interactions in higher education and instructional practices in K–12 schools. In an example of the former, Bettinger, Liu, and Loeb (2016) examined peer effects in college online classrooms by analyzing how peers interact with one another using rich student interaction data from online discussion forums. Exposure to more engaging peers increased students' probability of passing the course, earning a higher grade, and re-enrolling in the subsequent academic term. Another study by Aulck et al. (2019) examined how and why freshman seminars organized by interest group might have a positive influence on graduation and first-year retention rates. Using topic modeling to code more than 12,000 first-year interest group students' open-ended survey responses, they found that the social aspects of the seminars, particularly meeting new people and having friends and acquaintances in classes, were most frequently reported as the most valuable.

In an examination of microprocesses of teaching, meanwhile, Kelly, Olney, Donnelly, Nystrand, and D'Mello (2018) used both automatic speech recognition and machine learning to detect teachers' use of authentic questions, an important dimension of classroom discourse. Relatedly, Wang, Miller, and Cortina (2013) used an automated speech recognition tool to precisely classify the interaction patterns between teachers and students and provide timely feedback to teachers that could help them monitor students' active participation in classroom discussion. Although each of these two studies focused on only one dimension of teaching, Liu (2017) analyzed about 1,000 classroom transcripts and measured multiple teaching practices, including teacher–student turn-taking in classroom discussions, teachers' use of open-ended questions, and instructional routines. Some of these dimensions were found to consistently predict teachers' value-added scores to student achievement.

The aforementioned studies demonstrate the potential of using text-as-data methods in education research. In the current article, we illustrate a new application of these methods, capturing policy implementation and change processes in schools by analyzing school improvement planning and implementation reports. More important, we improve on prior studies that did not as thoroughly validate the measures derived from text analysis. Because automated text analysis requires researchers to regularly make key decisions and there is no universal standard to guide such decision making, text-as-data methods may generate unreliable or invalid measures (Grimmer & Stewart, 2013; Wilkerson & Casas, 2017). With the current study, we aim to show how researchers can use both substantive and statistical evidence to conduct comprehensive validation for measures derived from text analysis.

The Present Study

Policy Background of School Improvement for Underperforming Schools

This study explores reform strategies that underperforming schools planned and implemented to improve student achievement and reduce absenteeism, focusing on school reform efforts during the era of No Child Left Behind waivers and SIGs from 2010 to 2016. We focus on Washington State because this state largely adopted federal policy requirements and used three widely used policy instruments to turn around its underperforming schools: accountability and monitoring, funding/grants, and technical assistance to schools provided by improvement coaches (Hurlburt, Le Floch, Therriault, & Cole, 2011; Hurlburt, Therriault, & Le Floch, 2012).

During this period, Washington State implemented a multitiered identification and support system to remedy schools' underperformance. The state used three school improvement designations: *focus schools, priority schools*, and *SIG schools*. Focus schools were defined as those in the lowest 10% of subgroup performance based either on the 3-year average for subgroups on state assessments in English language arts and math (combined) or on an adjusted 5-year cohort graduation rate that was less than 60%. The state defined priority schools as those in the lowest 5% based on all students' performance across several criteria. The majority of schools identified had a 3-year average proficiency level for all students on state assessments in English language arts and math (combined) that was less than 40%, or in the lowest 5% based on the achievement index score,[1] or had an adjusted 5-year cohort graduation rate for all students that was less than 60%. SIG schools have to also be identified as priority schools. Other factors that could be considered when selecting SIG schools included geographic location, school size, and commitment and capacity to use SIG funds to substantially raise student achievement.

Once a school was identified as a focus, priority, or SIG school, typically that designation remained in place for 3 years. Schools would receive supplementary funding on top of their regular budgets; SIG schools were primarily funded through federal grants, whereas priority and focus schools were primarily funded through state funds. SIG schools were also required to follow federally prescribed school reform models. Almost all the SIG schools in Washington State adopted either the *transformation model* or the *turnaround model*. The transformation model requires replacing the principal, implementing curricular reform, and introducing teacher evaluations (based, in part, on student performance) into personnel decisions (e.g., rewards, promotions, retentions, and firing). The turnaround model includes all of the transformation model requirements, along with replacing at least 50% of the staff. Priority and focus schools received less, although still substantial, funding and assistance. Because they received state funds, they closely followed the state's guidelines on school turnaround, which are largely consistent with SIG models but have less strict requirements for replacing school personnel and tying educator evaluations to student growth. Overall, several features of these reform efforts—such as their relatively long duration, their systematic and dramatic approach to change, and the substantial influx of resources they prompted—make them a fertile ground for researchers and policymakers to learn useful lessons about school improvement strategies that move the needle for students.

To date, conventional evaluation studies of these school turnaround models have shown mixed effects on student achievement. Studies of SIG programs in California and Massachusetts using either regression discontinuity or difference-in-differences models found that the programs had positive effects on student achievement (Dee, 2012; Papay, 2015; Sun, Penner, & Loeb, 2017). However, a U.S. Department of Education study, using data from 22 states, found largely null impacts on test scores, high school graduation, and college enrollment for the cohort of SIG schools funded in 2010 (Dragoset et al., 2017). In several states that won Race to the Top funding or received No Child Left Behind waivers, research has yielded mixed evidence on the effectiveness of their school turnaround reforms.

Heissel and Ladd (2018) found negative effects from the programs in North Carolina, whereas Zimmer, Henry, and Kho (2017) found some positive effects in Tennessee, particularly among Innovation Zone schools that were governed and managed separately by three school districts. Two companion studies in Kentucky and Louisiana showed opposite findings: Over each of 3 years, Louisiana's focus school reforms had no measurable impact on school performance (Dee & Dizon-Ross, 2017), whereas Kentucky's focus school reforms led to substantial improvements in both math and reading achievement (Bonilla & Dee, 2017).

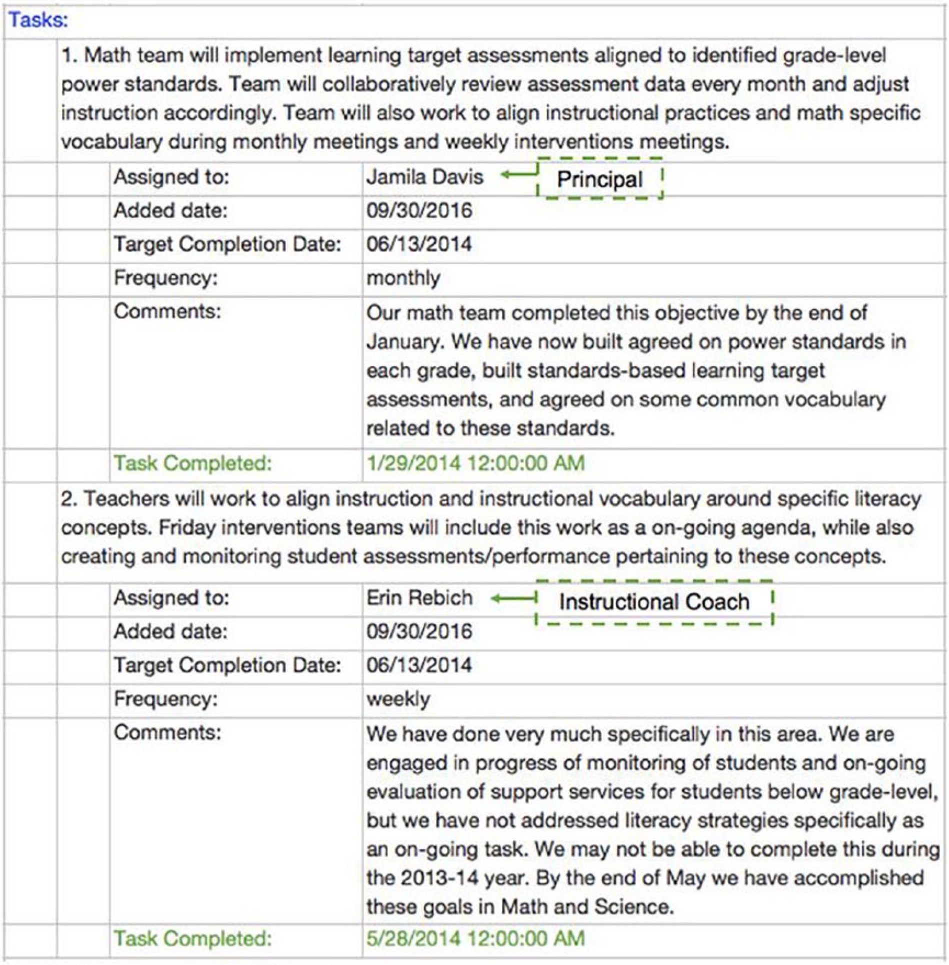
Some of the disparities in these results may be explained by sample selection and estimation strategies, as Guthrie and Henry's (2016) work in North Carolina illustrates. However, a more plausible explanation for the differences in findings across studies is the variation in the design and implementation of school reform interventions across schools, districts, and states (Dragoset et al., 2017). Given the state of the literature, it is apparent that another efficacy study using a "black box" approach would not be sufficient to inform future school improvement efforts. Rather, schools and districts need studies that use novel data and methods to investigate school improvement processes to generate actionable knowledge that can guide policy and practice directly.

Text Data

To develop a more detailed understanding of the mechanisms of change in schools, we analyze data on school reform processes collected through the Comprehensive School Improvement Planning and Implementation Reports (CSIPIRs) by Washington State's education agency, the Office of Superintendent of Public Instruction (OSPI). CSIPIRs are submitted by schools through a Web-based platform called Indistar. The state has specified seven principles of student and school success to guide a school's use of the reports: strong leadership; staff evaluation and professional development; expanded time for student learning and teacher collaboration; rigorous, aligned instruction; use of data for school improvement and instruction; safety, discipline, and social, emotional, and physical health; and family and community engagement. When using the Indistar system to build a school improvement plan, schools are required to select at least one indicator (from a bank of indicators provided by Indistar) for each of the seven principles. Selecting an indicator allows a school to see the evidence supporting that indicator, with the aim of providing evidence-based practices for schools.

The variation across schools then derives from the specific reform strategies that individual schools develop themselves, along with their implementation of these strategies. Once an indicator has been selected, a school is asked to describe and rate its current level of practice and to establish goals for what it will look like in practice if this indicator is fully achieved. Next, the school is asked to lay out specific tasks needed to achieve each goal, including designating the individual(s) responsible for the goal, the target completion date, and the frequency of the task. (See Figure 1 for an example.) A school is allowed to plan as many tasks as needed to achieve a goal. The school is also required to update its report periodically to mark the completion date for completed tasks, add comments on implementation, and explain how the school plans to sustain the task. This structured template helps schools to develop detailed information about their school improvement plans and implementation.

These reports provide a useful source of data on school improvement actions and activities for several reasons. They are not merely planning reports, but rather capture what schools actually implemented, as indicated by the date markers for task assignment and task completion. They are not merely compliance reports, either. In our interviews with personnel at 10 schools, many school leaders reported that because they could access the Indistar online tool anytime and anywhere, this reporting format was more convenient for coordination and communication among school staff than the old-fashioned paper format. Moreover, the Indistar system provides evidence associated with each indicator. School leaders indicated that they had developed more evidence-based planning and implementation with the Indistar system in place than they had before. The Indistar online tool and CSIPIRs were reported to help schools develop shared language and strategies among staff members. In addition, schools had little incentive to present lofty goals that they might be



Tasks:	
1. Math team will implement learning target assessments aligned to identified grade-level power standards. Team will collaboratively review assessment data every month and adjust instruction accordingly. Team will also work to align instructional practices and math specific vocabulary during monthly meetings and weekly interventions meetings.	
Assigned to:	Jamila Davis
Added date:	09/30/2016
Target Completion Date:	06/13/2014
Frequency:	monthly
Comments:	Our math team completed this objective by the end of January. We have now built agreed on power standards in each grade, built standards-based learning target assessments, and agreed on some common vocabulary related to these standards.
Task Completed:	1/29/2014 12:00:00 AM
2. Teachers will work to align instruction and instructional vocabulary around specific literacy concepts. Friday interventions teams will include this work as a on-going agenda, while also creating and monitoring student assessments/performance pertaining to these concepts.	
Assigned to:	Erin Rebich
Added date:	09/30/2016
Target Completion Date:	06/13/2014
Frequency:	weekly
Comments:	We have done very much specifically in this area. We are engaged in progress of monitoring of students and on-going evaluation of support services for students below grade-level, but we have not addressed literacy strategies specifically as an on-going task. We may not be able to complete this during the 2013-14 year. By the end of May we have accomplished these goals in Math and Science.
Task Completed:	5/28/2014 12:00:00 AM

FIGURE 1. *Examples of tasks written in the Comprehensive School Improvement Planning and Implementation Reports (CSIPIRs).*
Note. Staff names are pseudonyms.

unable to achieve later because the reports were submitted after schools had been identified for improvement and had already received federal or state funds, rather than being written during the grant competition stage. Furthermore, the state does not withhold funds, or hold schools accountable in other ways, for less ambitious plans or fewer tasks completed. The reporting is thus a nonconsequential requirement. Finally, the state provides coaches to identified schools to (a) support the development and implementation of improvement plans and (b) serve as a third-party monitoring mechanism. The above contexts provide us confidence in the validity of these data. As discussed later in the article, we also established our own procedures to further assess the quality of the text data using alternative data sources (e.g., interview data).

All identified schools in Washington State were required to submit CSIPIRs as of the 2011–2012 school year. We obtained these reports through a research–practice partnership with OSPI, then extracted information from these reports, from originals in PDF format into Excel spreadsheets using Python. Among the CSIPIR data that we received and cleaned (from 2011–2012 to 2015–2016), 55.2% of all unique tasks proposed were marked as completed with specific completion dates.[2] Incomplete tasks were either removed in later years' reporting or never marked with a completion date. Our subsequent analyses use only these completed tasks because they represent the completion of resource allocation and schools' committed actions. However, we acknowledge that our analyses using only completed tasks may cause upward bias in the results because schools may have abandoned strategies that they deemed were not yielding positive effects on student outcomes.

For some SIG schools, we used annual reports if they did not submit CSIPIRs, particularly in the early reform years (e.g., 2011–2012). Only SIG schools were required to submit annual reports, which have structured reporting elements similar to those in CSIPIRs but were submitted only at the end of the school year and included a summary of the year's completed initiatives. SIG annual reports were used by the state school improvement coaches as part of their validation process of CSIPIR data. All tasks mentioned in the annual reports would be counted as completed tasks, per the requirement of the reports. Because schools are asked to submit three CSIPIRs per year, the total number of reports yielded from these schools by the end of the 2015–2016 school year is 2,873 CSIPIRs and 85 SIG annual reports[3] (in the early years, some schools submitted two CSIPIRs per year). We have 25,486 completed tasks from CSIPIRs and 510 tasks from SIG annual reports. Texts with less than seven tokens were removed (about 5.8% of the total sample), however, because they provide little useful information on what schools actually did and are not suitable for the LDA model. After trimming those from the corpus, we have 23,997 unique tasks from CSIPIRs and 502 tasks from SIG reports. The next section describes our approach to deriving quantitative measures of reform strategies from these data.

Text Analysis

Schools conceived of their own reform strategies and used their own words to report the tasks that they undertook and completed. Our goal of text analysis is to identify schools' fine-grained reform strategies as well as the extent to which they were implemented in each school using this large volume of unstructured textual data. After carefully considering many text-as-data methods, including dictionary methods, clustering, and supervised methods, the topic modeling approach, specifically LDA (see Blei, 2012; Blei, Ng, & Jordan, 2003), stands out as the most appropriate one. Rather than requiring researchers to condition on known constructs or topics beforehand, LDA uses modeling assumptions and properties of texts to generate a set of topics and simultaneously assign tasks to those topics. It is particularly useful when learning the patterns of text data or trying to identify topics that are theoretically meaningful but perhaps understudied or previously unknown. Using LDA, we are able to condense thousands of diverse CSIPIR text entries into a limited number of discrete and sensible categories, or topics, and simultaneously derive the composition of topics for each text entry.

LDA is a generative statistical model that identifies the latent topics and corresponding proportions that compose a document. LDA assumes that each document (a reform task in our setting) is a mixture of topics. For each task, π_{ik} represents the proportion of task i dedicated to topic k. Each task collects the proportions across topics, as $\pi_i = \left(\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK}\right)$. We used an R package (-stm-) to implement the analysis.

LDA allows us to estimate both topic prevalence (e.g., the proportion of a task discussing each topic) and latent content constructs with observed information about school improvement processes. In contrast to other clustering methods that assign documents to only one topic (or latent construct), LDA analysis aims to discover the latent topics across all tasks and represents a given task as a set of topic weights, rather than assigning each task to a single topic. The topic weights, as indicated by the proportion of texts, can then be aggregated across all task entries for a given school in a given year to produce an overall assessment of task emphasis. The topic proportion indicates the prevalence of reform strategies in schools and reflects a mix of factors, such as the time that schools spent on a reform topic, the importance of the reform topics, or the depth that a school engaged in this reform topic. Our measure is similar to that used in a U.S. Department of Education study of the proportion of practices under SIG reform topic areas (Dragoset et al., 2017) collected via surveys, but with greater accuracy and comprehensiveness. Below, we describe how we processed the raw text data, derived a longitudinal measure of topic prevalence at the school-year level, and used alternative data sources (such as human-coded metrics and interview data) to assess the validity and robustness of the results.

Preprocessing. The first step was to define the text features to be modeled using LDA. A standard practice is to exclude common "stop words" (such as "the" or "and") and stem words that have the same root meaning (e.g., "learning" becomes "learn"). We also reviewed word lists to identify and include domain-specific phrases (e.g., "professional learning communities" [PLC]) and to group references in the same "named entity" (such as "professional learning communities" and the acronym "PLC") using a 3-gram approach.

Topic Analysis Using LDA. The topic estimation was conducted at the individual task level. We only used unique entries so that tasks carried over from 1 year to the next would not create duplicate information that might distort our topic extraction. To capture the cumulative nature of longer term tasks, we then aggregate tasks to the school-year level by accounting for the number of years each task was mentioned in the reports.

Topic Aggregation. Then, to aggregate the data from task level to school-year level, we weighted each individual task by the proportion of the wording of the task out of all unique tasks in and up to that year, then summed the weighted topic proportions across all tasks. This differential weighting is based on several reasons. First, we observed that if a school gave a higher priority to the task, the school would use more words to provide more specific and concrete information about the task. This observation is based on anecdotal evidence from conversations with principals and state-assigned coaches, as well as from our manual reading of many tasks written in the reports and comparisons among tasks written by the same narrator. Second, the number of words a school devoted to describing a task can also be reviewed as a precision weighting in linguistic analyses. Topic proportion allocations to topics are often less precisely estimated for tasks written with fewer words in topic analyses. Third, we did estimate the relationships between reform topics and school performance (e.g., school average achievement in math and reading, and school average absenteeism) without weighting on the number of words (see Supplemental Tables S4 and S5, available in the online version of the journal). The results are largely similar to the findings using the weighted measures in Tables 5 and 6. However, the coefficients of weighted measures in Tables 5 and 6 are more efficiently estimated than the coefficient estimates of the unweighted measures, as evidenced by the smaller standard errors. We thus prefer the measures weighted by the number of words of tasks.

Moreover, because these school reform efforts are dramatic, fundamental, and continuous, they often involve tasks that are long term and aim to build schools' basic capacity, such as providing teachers with professional development, building leadership teams in schools, and engaging parents and communities. Prior studies have observed stronger cumulative effects of these types of reform strategies on student achievement than year-to-year effects (May & Supovitz, 2006; Sun et al., 2017). The cumulative proportion here aims to capture this nature of the reform efforts, as illustrated in the following equation:

$$\mathbf{p}_c = \sum\nolimits_{k=1}^{k=n} p_{k,c} * w_k ,$$

where $\mathbf{p}_c$ is the proportion for topic c at the school-year level, $p_{k,c}$ is the proportion of task k on this topic, and w_k is the proportion of words in task k out of the total number of words in all unique tasks in and up to that school year. We then sum across all tasks loaded onto topic c. For example, if a task appears in a document in the second year of reform for a school, w_k is calculated using the number of words in all the tasks articulated in the first 2 years of reform. Thus, $\mathbf{p}_c$ is calculated as the cumulative task proportion on topic c in the first 2 years of reform. This calculation is designed to better capture the totality of a school's emphasis up until that time.

Validation. Validation is essential for automated text analysis methods such as LDA because the researcher makes design decisions that have important implications for the findings. Validation needs to combine both statistical tools and careful human judgment. To make sure computer-generated topics indeed capture the "true" topic in the text, we ran a number of models by specifying the number of topics to arrive at, ranging from 10 to 30 topics. Although the -stm- package provides several statistical indices to indicate model fitness, the "best" model needs to capture the topics of interest to the researcher (Roberts et al., 2014; Wang, Paisley, & Blei, 2011). As a result, model choice is typically based at least partially on subjective considerations similar to those in more traditional qualitative research (Grimmer & Stewart, 2013; Saldana, 2009). In this study, we first used several model diagnostic statistics (such as semantic coherence and exclusivity) that pointed to either a 15-topic model or a 20-topic model as the best fit. We also asked human coders to assess whether tasks loaded highly on a given topic indicated coherent meaning (as discussed further below). This subjective evaluation led to the conclusion that the 20-topic model was optimal. In the "Results" section, we illustrate the process and results for establishing content validity or semantic coherence, internal structure, and relationships to other variables, including predictive validity, per American Educational Research Association (AERA)/National Council on Measurement in Education (NCME)/American Psychological Association (APA) test standards (Chan, 2014).

Sample and Structured Administrative Data

We then linked these reform process measures with school contextual and student outcome measures from state administrative data sets to examine (a) which schools and communities adopted which types of reform strategies and (b) how the reform processes explain the variation in the effects of school improvement efforts on student outcomes. We used both student absenteeism and achievement on state standardized tests to measure school improvement outcomes. OSPI collects data on four types of absences: full-day excused, part-day excused, full-day unexcused, and part-day unexcused. A *full-day absence* is defined as missing more than 50% of instructional time during a day. In our analyses, we combine excused and unexcused absences because such division can be imprecise if students, parents, or schools treat them as fungible. Along with running analyses on the raw numbers of partial and full days students missed, we created a chronic absenteeism measure for students who were absent for 15 or more full school days.[4]

TABLE 1

Number of Treatment Schools That Have Test Score Data by Reform Type and Cohort

School year	SIG	Priority	Focus
2015–2016	0	68	116
2014–2015	0	56	131
2013–2014	10	34	79
2012–2013	24	15	66
2011–2012	24	0	0
2010–2011	0	0	0

Note. This table includes the number of schools that have text data and one outcome measure (e.g., either achievement or chronic absenteeism). Our analysis only includes the first designation of a given school. SIG = School Improvement Grant.

Achievement on state standardized test scores is standardized within a given grade, year, and test to account for differences in tests across grade levels, subjects, and years. Tests include Smarter Balanced Assessments in math and English language arts in Grades 3 to 8 and 11, Washington State's Measurements of Student Progress tests, and end-of-course exams in Grades 9 to 12, among others. If a student took more than one math test in a year (e.g., geometry and algebra), we took the average of the standardized scaled scores as the measure for this student.

Table 1 summarizes the number of schools identified as SIG, priority, and focus schools for which we have both text data and student outcome measures in either absenteeism or achievement during school years 2010–2011 through 2015–2016. In 2010–2011, 18 SIG schools were identified; however, because the state did not adopt the Indistar system until 2011–2012, we do not have their CSIPIRs or SIG annual reports. Although in 2011–2012 the state identified 28 SIG schools, we are missing either the reports or the student outcome measures for four of those schools, so there are 24 SIG schools in our analytic sample. Since 2012–2013, more priority and focus schools were included in the analysis over time. In total, our sample includes 318 schools and 623 school-year observations. As shown in the online Supplemental Table S1, the final analytic sample is representative of all identified schools in terms of prereform characteristics and performance.

Table 2 summarizes the characteristics and performance of these identified schools as well as

TABLE 2

School-Level Characteristics and Performance by Reform Type

	SIG	Priority	Focus	Nonreform
% White	0.27	0.39	0 .39	0.56
	(0.26)	(0.32)	(0.25)	(0.29)
% African American	0.13	0.06	0.05	0.04
	(0.19)	(0.11)	(0.09)	(0.07)
% Hispanic	0.36	0.34	0.42	0.16
	(0.32)	(0.31)	(0.26)	(0.19)
% Asian	0.06	0.03	0.04	0.05
	(0.09)	(0.06)	(0.07)	(0.08)
% Pacific Islander	0.01	0.01	0.01	0.01
	(0.02)	(0.02)	(0.02)	(0.02)
% Native American	0.10	0.11	0.03	0.02
	(0.2)	(0.22)	(0.08)	(0.08)
% Multiracial	0.05	0.06	0.05	0.05
	(0.05)	(0.05)	(0.04)	(0.05)
% Eligible for free or reduced-price lunch	0.78	0.71	0.69	0.41
	(0.19)	(0.26)	(0.17)	(0.27)
% English language learner	0.36	0.29	0.36	0.17
	(0.26)	(0.28)	(0.23)	(0.19)
% Homeless	0.05	0.06	0.05	0.03
	(0.03)	(0.07)	(0.06)	−0.04
% Special education	0.13	0.15	0.14	0.16
	(0.05)	(0.11)	(0.05)	(0.21)
Prior student academic achievement	−0.6	−0.57	−0.33	N/A
	(0.19)	(0.30)	(0.23)	
Student academic achievement	−0.42	−0.47	−0.31	−0.05
	(0.25)	(0.37)	(0.29)	(0.44)
Full-day absences	12.66	9.94	9.89	7.18
	(5.85)	(6.12)	(4.67)	(5.32)
Part-day absences	5.83	4.09	4.81	2.89
	(7.86)	(5.37)	(5.85)	(4.59)
% Chronic absenteeism—full day	0.41	0.33	0.33	0.24
	(0.16)	(0.17)	(0.12)	(0.17)
N (school year)	86	200	486	17,404

Note. The data were from 2010 to 2016. This table includes schools' first reform type identifications. Nonreform schools did not have prior student outcome measures per definition because they did not have a reform start date. The sample sizes reported here only reflect the analytic sample that provides all the demographic information. The sample sizes for the absence measures are 39, 200, 485, and 9,465 for SIG, priority, focus, and nonreform schools, respectively. Standard deviations are reported in parentheses. SIG = School Improvement Grant.

nonreform schools in the state. Identified schools, on average, serve larger proportions of historically underserved students—including students of color, low-income students, and homeless students—than do nonidentified schools. Students in identified schools are also relatively lower achieving and more likely to be chronically absent.

Results

Model Diagnostics Statistics

The LDA approach requires researchers to specify the number of topics. OSPI specified seven principles of school improvement; Bryk, Sebring, Allensworth, Luppescu, and Easton

(2010) also identified five essential supports for school improvement. These categories of school improvement efforts are broad (such as “building school leadership teams” or “developing teacher capacity”) and do not discuss specific strategies schools might employ. Aiming to discover new and more specific reform strategies, we began the modeling process by specifying 10 to 30 topics, and then, we used diagnostic statistics to aid our model selection, as illustrated in Figure 2.

The first diagnostic statistic is *semantic coherence*, or the degree to which words are internally consistent. In Figure 2a, the *y* axis indicates log probabilities. Large negative values indicate that top words do not co-occur often, whereas values closer to zero indicate that top words tend to co-occur more often. In our case, the 10-topic model has the highest semantic coherence, whereas the 15- and 20-topic models are slightly worse and the 25- and 30-topic models substantially worse.

The second diagnostic statistic we use, *exclusivity*, summarizes the harmonic mean of the probability of all the top words under a topic and the exclusivity to that topic (Roberts et al., 2014). The bigger the value on the *y* axis of Figure 2b, the better the model performs in terms of separating one topic from the others. In our case, the 20-topic model is better than the 15-topic model, and both of these are much better than the 10-topic model.

Given that a topic that is both cohesive and exclusive is more likely to be semantically useful (Roberts et al., 2014), the 15- and 20-topic models appear to provide better balance between semantic coherence and exclusivity than the 10-, 25-, or 30-topic model. These statistics are helpful only to the extent that they provide us with general guidance on model selection. The coherence and exclusivity of the overall model do not directly indicate whether each topic of the model represents a conceptually and practically meaningful “theme.” To assess this, we need human coders to further evaluate the content validity of the topics.

Content Validity Check

Content validity denotes the extent to which our topics identify coherent sets of tasks and measure conceptually sound constructs. We used two expert coders on our research team to assess the content validity of our topics.[5] Having expert coders who have both practical and theoretical knowledge is critical to assess topics’ content validity. We then developed a rubric to rate topic coherence and the extent to which the coherent topic is practically meaningful and consistent with the literature (included in the online Supplemental Table S3; Mimno, Wallach, Talley, Leenders, & McCallum, 2011). Using a scale of 1 to 4, the two experts coded topics from the 15-, 20-, and 30-topic models. Interrater reliability, as measured by Krippendorff’s alpha, ranged from .81 to .89, depending on the individual topic model. The two coders first independently labeled each topic and rated its coherence by reading a sample of tasks with the highest loadings on a given topic. After going through this process for each model specification, the two coders compared and discussed their ratings. Most of the differences were only a 1-point difference (e.g., one coder rated a topic’s coherence as a 2, whereas the other rated it a 3). If the coders could reach agreement, they adjusted their individual scores to the score they agreed to. If they could not, they preserved their original ratings (for only one topic did the two coders preserve their own original ratings).

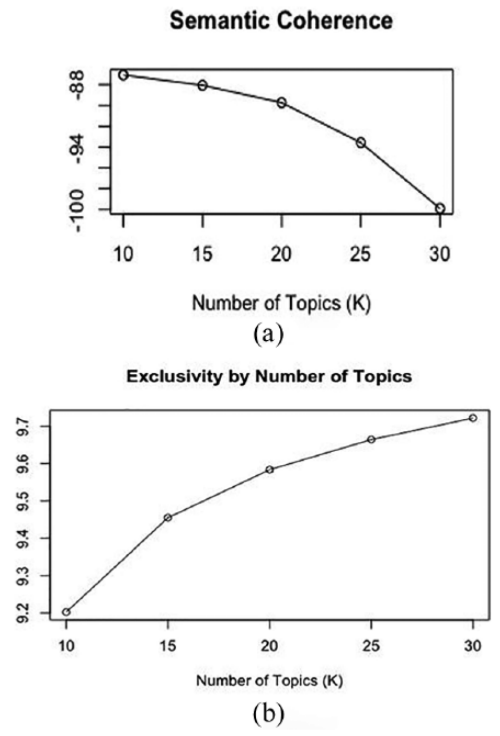


FIGURE 2. *Topic model diagnostic statistics.*

The results of the rating process indicated that the 20-topic model was optimal. Besides having a higher average coherence rating than the 15-topic model ($\mu_{Model\_20} = 3.18$, $\mu_{Model\_15} = 2.9$), 75% of the topics in the 20-topic model have a rating of 3 or higher, whereas 73% of ratings in the 15-topic model are rated 3 or higher. More significantly, 45% of topics in the 20-topic model have a rating of 4, compared with only 13% of topics in the 15-topic model. Because an average coherence rating of lower than 3 for a given topic casts doubt on its content validity per definitions of these categories in our rubric descriptions, only topics with an average rating of 3 or higher were used for subsequent analysis. The second column of Table 3 provides the average ratings from the two coders for each topic of the 20-topic model.

Reform Strategy Prevalence

The last column of Table 3 indicates that topics vary in prevalence, as indicated by the means of average topic proportions at the school-year level.[6] The topic with the highest mean proportion of 0.095 signifies that schools on average spent 9.5% of their reform efforts annually on Topic 11 ("building leadership teams to set goals and review data for school improvement"). Other topics with high mean proportions include Topic 1 ("interventions and supports for promoting positive student behaviors"; $\mu = .70$, *SD* = 0.068), Topic 3 ("engaging parents in student academic and behavioral learning in school"; $\mu = .084$, *SD* = 0.078), and Topic 12 ("teacher instructional improvement via walkthroughs, observations, and feedback"; $\mu = .075$, *SD* = 0.081). Topics with low mean proportions include Topic 10 ("administering common assessments and disaggregating data to differentiate interventions"; $\mu = .021$, *SD* = 0.039) and most of the topics with low coherence ratings. Moreover, we observe large variations for many topics, with standard deviations equal to 2 times as large as the mean. This shows that the topic prevalence or prioritized reform strategies varied across schools and over time. We further explore this variation in the next section.

Variations in Reform Strategies by SIG, Priority, and Focus Schools Over Time

As shown in Figure 3, SIG schools had more changes in the proportions of tasks implemented over time than priority and focus schools did. For example, from Year 1 to Year 3, SIG schools greatly increased the implementation of Topic 8 ("extending instructional time and aligning curriculum or assessments to standards") and Topic 15 ("setting goals for and recognizing teachers and students' growth"). In contrast, the topic proportions for priority and focus schools were relatively stable over time. The patterns for priority and focus schools are largely similar to one another. Compared with SIG schools, they seem to implement more tasks on Topic 1 ("interventions and supports to promote student behaviors"), Topic 11 ("leadership teams setting goals and reviewing data for school improvement"), and Topic 12 ("teacher instructional improvement via walkthroughs, observations, and feedback").

These patterns are understandable, considering the multitiered accountability and support system Washington State implemented. SIG schools followed federal guidelines that particularly emphasized strategies of aligning curriculum with assessments and standards and extending instructional time, as well as strategies of promoting students' growth and rewarding teacher performance based on student growth. Priority and focus schools were funded through state resources and were encouraged but not required to follow the SIG guidelines. These schools used the Indistar system to align their efforts with the seven state principles of school reform. Therefore, the following activities stand out at priority and focus schools when compared with SIG schools: building a strong leadership team, implementing new evaluation systems for teachers and principals, and developing positive student behaviors for social–emotional learning and a safe school climate. In addition, SIG schools received stronger treatment over time due to increased accountability pressure, whereas priority and focus schools did not experience the same kind of pressure. This consistency between reform strategies and reform type further sheds light on the promise of the text analysis results.

Internal Structure of Reform Strategy Measures

We next examined the *internal structure* of the topics—in this case, investigating how topics that theoretically should be related are in fact related, or how topics that theoretically should not be related are in fact unrelated. For example, as

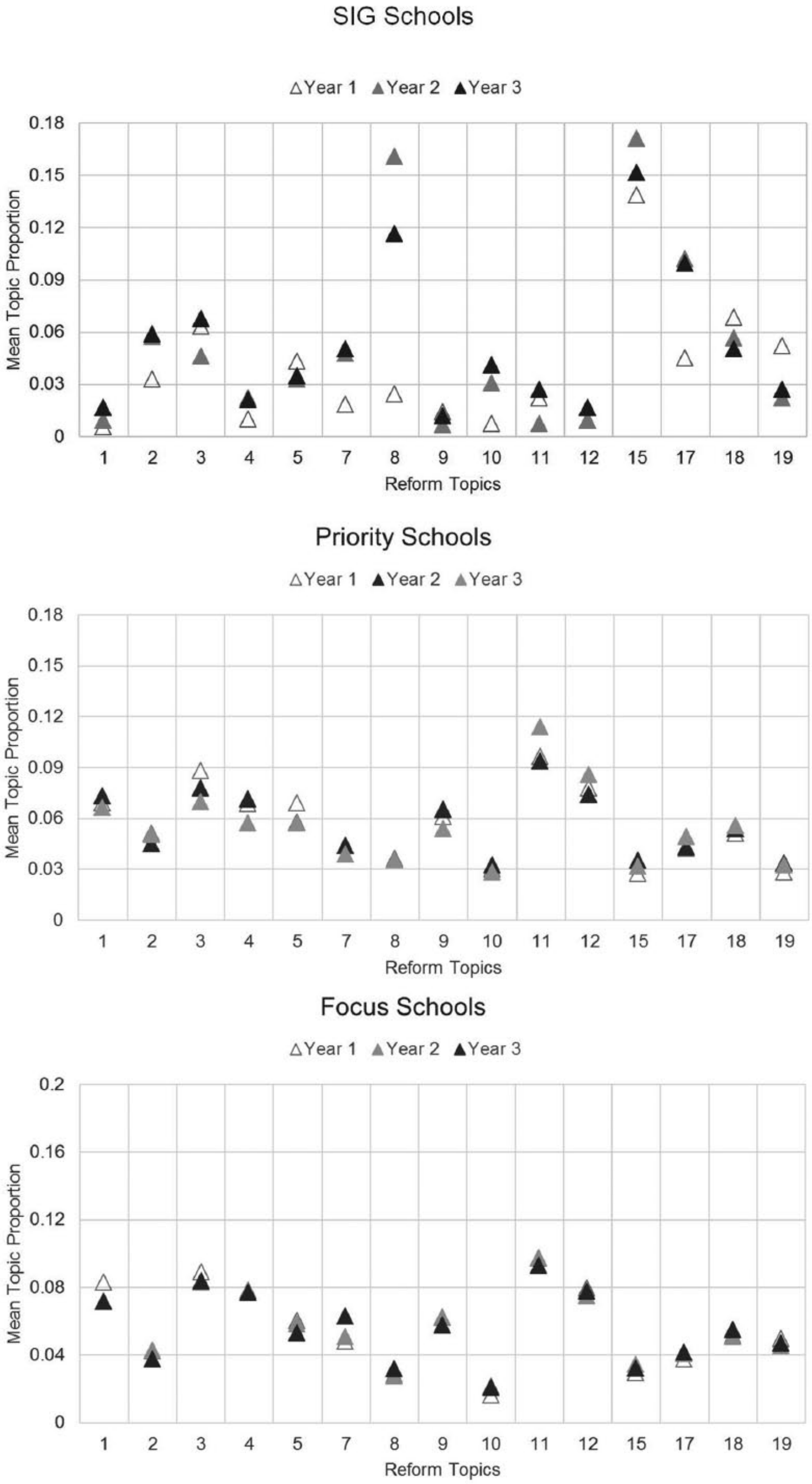
TABLE 3

Descriptives of Reform Topics

Reform topics	*M* coherence rating	*M* (*SD*)
Topic 11. Leadership teams setting goals and reviewing data for school improvement	4	0.095 (0.082)
Topic 3. Engaging parents about student academic and behavioral learning in schools	4	0.084 (0.078)
Topic 12. Teacher instructional improvement via walkthroughs, observations, and feedback	4	0.075 (0.081)
Topic 1. Interventions and supports for promoting positive student behaviors	4	0.070 (0.068)
Topic 4. Planning, providing, and evaluating professional development for instructional improvement	4	0.068 (0.059)
Topic 9. Teacher team (e.g., grade-level team, PLC) activities (e.g., reviewing data, planning, aligning standards, developing interventions)	4	0.061 (0.077)
Topic 5. Monitoring student progress and using data to develop interventions	4	0.057 (0.058)
Topic 18. Collecting, analyzing, and aligning student assessments	4	0.052 (0.051)
Topic 7. Using assessment data to identify students for targeted support	3	0.047 (0.06)
Topic 2. General parent and community outreach	4	0.045 (0.055)
Topic 17. Extending learning time (or opportunities) for students and staff	3	0.044 (0.06)
Topic 20. (Low coherence)	2	0.043 (0.063)
Topic 16. (Low coherence)	2	0.039 (0.048)
Topic 19. Improving special education	3.5	0.038 (0.049)
Topic 15. Setting goals for and recognizing teachers' and students' growth	3	0.038 (0.071)
Topic 8. Extending instructional time and aligning curriculum or assessments to standards	3	0.038 (0.074)
Topic 13. (Low coherence)	2	0.037 (0.046)
Topic 14. (Incoherent)	1	0.028 (0.042)
Topic 10. Administering common assessments and disaggregating data to differentiate interventions	3	0.021 (0.039)
Topic 6. (Low coherence)	2	0.021 (0.054)

Note. Topic proportions are at the school-year level; table is sorted by mean cumulative topic proportions. PLC = professional learning communities.

shown in Table 4, Topic 1 ("interventions and supports for promoting positive student behaviors") and Topic 12 ("teacher instructional improvement via walkthroughs, observations, and feedback") have a near-zero correlation (ρ = −.006) because these two reform strategies do not necessarily



SIG Schools

Priority Schools

Focus Schools

(continued)

FIGURE 3. *Topic proportions by reform types and over reform years.*
Note. Reform topic labels:
Topic 1. Interventions and supports for promoting positive student behaviors
Topic 2. General parent and community outreach
Topic 3. Engaging parents about student academic and behavioral learning in schools
Topic 4. Planning, providing, and evaluating professional development for instructional improvement
Topic 5. Monitoring student progress and using data to develop interventions
Topic 7. Using assessment data to identify students for targeted support
Topic 8. Extending instructional time and aligning curriculum or assessments to standards
Topic 9. Teacher team (e.g., grade-level team, PLC) activities (e.g., reviewing data, planning, aligning standards, developing interventions)
Topic 10. Administering common assessments and disaggregating data to differentiate interventions
Topic 11. Leadership teams setting goals and reviewing data for school improvement
Topic 12. Teacher instructional improvement via walkthroughs, observations, and feedback
Topic 15. Setting goals for and recognizing teachers' and students' growth
Topic 17. Extending learning time (or opportunities) for students and staff
Topic 18. Collecting, analyzing, and aligning student assessments
Topic 19. Improving special education.
SIG = School Improvement Grant; PLC = professional learning communities.

depend on one another, with one focusing on student behaviors and the other focusing on teacher evaluation. In contrast, Topic 12 and Topic 15 ("setting goals for and recognizing teachers' and students' growth") are significantly and negatively correlated ($\rho = -.246, p < .001$). Although both Topic 12 and Topic 15 describe strategies targeting teachers—which explains why they are correlated—schools that spent more resources and time on Topic 12 often spent less on Topic 15. Our interviews with school leaders suggest they viewed Topic 12 as a strategy for supporting teachers' professional growth and Topic 15 as an incentive-driven strategy. Moreover, Topic 12 summarizes a process-oriented reform strategy, whereas Topic 15 summarizes an outcome-oriented reform strategy. In contrast, Topic 15 is significantly and positively correlated with Topic 8 ("extending instructional time and aligning curriculum or assessments to standards"; $\rho = .162$, $p < .001$), which makes sense because both of these topics focus on supporting student and teacher learning and tie learning processes together with learning goals and standards.

Further Validation of the Reform Strategy Measures Using a Different Data Source

Another form of validity evidence is correlation with another measure of the same or a similar construct gathered from a different data source. To further validate our reform strategy measures, we interviewed a number of principals and other staff members from 10 schools about 3 to 6 months after the schools submitted their reports. If the schools had fabricated the reports, school staff would have had difficulty recalling their content months after submission. The 10 schools varied in student population, educational level, reform type (SIG, priority, and focus), and geographic location, as well as in their student achievement gains up to the year in which they were interviewed. We asked interviewees to freely describe the important initiatives they undertook in the last school year to transform their schools.

About 82% of the 10 most prevalent topics in the schools' reports were mentioned as top initiatives by school administrators. In 4 of the 10 schools, the principals and staff referenced nine or 10 of the top 10 topics in the reports, and in the other 6 schools, staff mentioned seven or eight of the top 10 topics. In particular, among these 10 interviewed schools, three of them were SIG schools who submitted SIG annual reports 1 to 4 years previous to the interview. Twenty of 23 of the top areas identified from interviews were also found in SIG annual reports, an alignment of 87%. The high alignment between reports and interviews serve as a robustness check for the LDA priority measures, suggesting that the text analysis results are similar to those derived from interview data, with the additional advantage of being feasible to obtain on a much larger scale and with relatively low cost.

The Predictive Validity of Reform Strategy Measures

If the quantitative measures derived from the topic modeling represent meaningful distributions

TABLE 4

Pairwise Correlations Among Reform Strategy Topics

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 15	Topic 17	Topic 18
Topic 2	−.095													
Topic 3	−.020	.024												
Topic 4	.052	−.134	−.104											
Topic 5	−.045	−.106	−.043	−.015										
Topic 7	−.009	.091	−.088	−.018	−.081									
Topic 8	−.150	−.012	−.143	−.137	−.107	−.018								
Topic 9	−.010	−.130	−.102	.024	−.119	−.049	−.099							
Topic 10	−.109	−.051	.002	−.010	−.081	−.065	−.032	−.007						
Topic 11	−.012	−.138	−.097	.089	−.021	−.099	−.125	.027	−.094					
Topic 12	−.006	−.129	−.007	.001	−.042	−.040	−.148	−.007	−.084	.086				
Topic 15	−.194	0.015	−.144	−.202	−.129	−.109	.162	−.109	.079	−.237	−.246			
Topic 17	−.063	−.020	−.095	−.142	.023	−.062	−.037	−.094	−.051	−.138	−.170	.133		
Topic 18	−.117	.015	−.138	−.005	−.104	−.010	−.035	−.108	.131	.068	−.098	−.099	−.081	
Topic 19	−.031	.018	−.101	.048	.013	−.010	−.164	−.007	−.101	−.081	−.009	.023	−.091	−.087

Note. The white background indicates positive correlation coefficients whereas the light gray indicates negative correlation coefficients.
Reform topic labels:
Topic 1. Interventions and supports for promoting positive student behaviors
Topic 2. General parent and community outreach
Topic 3. Engaging parents about student academic and behavioral learning in schools
Topic 4. Planning, providing, and evaluating professional development for instructional improvement
Topic 5. Monitoring student progress and using data to develop interventions
Topic 7. Using assessment data to identify students for targeted support
Topic 8. Extending instructional time and aligning curriculum or assessments to standards
Topic 9. Teacher team (e.g., grade-level team, PLC) activities (e.g., reviewing data, planning, aligning standards, developing interventions)
Topic 10. Administering common assessments and disaggregating data to differentiate interventions
Topic 11. Leadership teams setting goals and reviewing data for school improvement
Topic 12. Teacher instructional improvement via walkthroughs, observations, and feedback
Topic 15. Setting goals for and recognizing teachers' and students' growth
Topic 17. Extending learning time (or opportunities) for students and staff
Topic 18. Collecting, analyzing, and aligning student assessments
Topic 19. Improving special education.
PLC = professional learning communities.

of schools' reform strategies, these measures should have some power to predict changes in student absenteeism and achievement.

Student Absenteeism. We first examined the relationship between the topic proportions and student attendance, as attendance has a positive and statistically significant relationship with academic achievement (e.g., grade point average and standardized test scores in reading and math) for both elementary and middle school students (Gottfried, 2010). Poor attendance has serious implications for later outcomes. Prior research has found that students who eventually dropped out of high school missed significantly more school days in the first grade than their peers who graduated from high school did. In eighth grade, this pattern was even more apparent, and by ninth grade, attendance was shown to be a key indicator significantly correlated with high school graduation (Allensworth & Easton, 2005; Hickman, Bartholomew, & Mathwig, 2007).

We used topic prevalence at the school-year level to predict a school's current-year absences on each of these three measures: full-day absence, partial-day absence, and chronic absenteeism rate. These three measures all have their own strengths. Full-day absence is the most widely used measure in prior work as it is often the measure available. Recent research suggests that partial-day absence can account for at least half of total absences in secondary schools and can serve as a better measure on student engagement (Whitney & Liu, 2017). Chronic absenteeism rate is used in school accountability systems under Every Student Succeeds Act (ESSA) and is most useful for policymakers. Thus, these three measures complement each other and provide a more comprehensive assessment of the ways in which the identified topics associate with student absenteeism. We regressed each of the absence measures on each of the 15 coherent reform topics separately. Schools that were simultaneously implementing more tasks may have had to distribute resources and energy thinly, which may have affected their successful implementation of any one task. Our models further account for the number of tasks that schools were implementing in a given year.[7] Our models also control for pre-reform achievement level[8] and school characteristics (e.g., percentages of students eligible for free- or reduced-price lunch, English language learner [ELL] students, homeless students, historically underserved students of color [Hispanic/Latinx, African American, Native American or Alaskan Native, Asian Pacific Islander, and multiracial], and students with disabilities). Because the analysis includes multiple observations for individual schools over time, we clustered the standard errors at the school level.

Both Topic 9 ("teacher team activities [e.g., reviewing data, planning, aligning standards, developing interventions]") and Topic 10 ("administering common assessments and disaggregating data to differentiate interventions") are significantly negatively correlated with both the average days of full-day and part-day absences and the rate of chronic full-day absences. For example, a 10 percentage point increase in Topic 9 is associated with a reduction of 1.03 full-day absences, 1.68 part-day absences, and 2.41 percentage points of chronic absence rate. A 10 percentage point increase in Topic 10 is associated with a reduction of 1.64 full-day absences, 1.41 part-day absences, and 4 percentage points of chronic absence rate.[9] Although we cannot interpret these coefficients causally, it is useful to benchmark these coefficients using other related studies. In a recent intervention study that provides parents' information on their children's missed school days and misbeliefs about the importance of regular school attendance, treated students show a reduction of 0.5 full-day absences and 1.4 percentage points of chronic absenteeism rate (Robinson, Lee, Dearing, & Rogers, 2018). Our coefficients are about twice as large as those in Robinson et al.'s intervention. Although it is helpful to contextualize the size of the coefficients in our study by comparing them with the effect sizes in prior studies, our study is descriptive in nature. Given potential omitted variables might bias our results, the coefficients can only be interpreted as associations.

These relationships can be explained by the nature of the activities the reform topics entail, such as communicating student data with families and teachers and developing targeted interventions based on student needs. As illustrated by the exemplary tasks pertaining to Topic 9 below, effective teacher team activities include developing teachers' capacity to use student data, as well as centering team activities around student learning and adopting targeted interventions for at-risk students (Lachat & Smith, 2005). Prior studies of programs that include these elements of monitoring student attendance, suspensions, assessment, and course grades to provide individualized attentions to at-risk students showed positive effects in small randomized control trial studies (e.g., Sinclair, Christenson, Evelo, & Hurley, 1998; Sinclair, Christenson, & Thurlow, 2005).

> Topic 9, Task 1: Department and grade level planning notes will be submitted on a monthly basis to principal who will review and give needed feedback and support. Data will also be shared at these meetings related to assessments, behavior, grades, and attendance to best support students.

> Topic 9, Task 2: We will do Benchmark testing on students 3x times a year, in DIBELS NEXT. Then we will progress monitor intensive and strategic students 2x a month. We will look at data in grade level teams to brainstorm strategies to help struggling students. Grade level teams will decide on which students need additional interventions and monitor their progress.

> Topic 9, Task 3: All grade levels create SMART goals for mathematics to align with the SBA (Smarter Balanced Assessments). The students who have not met the targeted standard will receive intentional instruction in this area, until the next assessment period. Some grade levels have overlapping SMART goals to ensure all students are making progress. This also helps students maintain their learning and move on at the same time.

Topic 10 ("administering common assessments and disaggregating data to differentiate interventions") depicts a set of practices of educators analyzing a variety of student achievement and growth data to adjust their instructional decisions. These instructional decisions include grouping students so that teachers can provide targeted supports, or reteaching certain materials. In other words, these activities are similar to Topic 9 in terms of educators making data use as part of ongoing routines, but with greater emphasis on collecting and analyzing assessment data. Although there are limited causal studies of how these practices influence student attendance, prior research does shed some light on the promise of these ongoing data use to improve attendance (Balfanz & Byrnes, 2013). Particularly, a recent multisite randomized control trial aimed at improving teachers' use of student data revealed a positive and significant impact on teachers' self-reported positive relationships with students (Borman, Bos, O'Brien, Park, & Liu, 2017). These positive relationships may enable educators to more effectively work with students to overcome their challenges and help attract students to attend classes.

> Topic 10, Task 1: Literacy Data collected through Fountas and Pinnell assessments, spelling inventories, and Scholastic reports are used to organize students for small group instruction in reading essential classes. Students are regrouped based on changes in performances. Pre and Post unit assessments are used to measure students' growth within a unit based on the district frameworks. Measures of adequate progress data is used to understand the growth patterns of specific classrooms and inform classroom instruction. Math Data from common Pre and Post unit assessments is used to re-organize students into groups for pre-teaching in the math essentials classes. Results from the state MBAs (Math Balanced Assessments) are used to inform regrouping students for re-teaching opportunities. Students are re-assessed after several weeks of re-teaching. Teachers grade assessments together. We have attached examples of how students are grouped and organized for small group instruction and examples of how data is represented for use in department meetings.

> Topic 10, Task 2: All students who have scored below standard on the Spring Benchmark Assessment (grades K-2) and the Spring state summative assessment (grades 3 and 4) will be assessed and placed in appropriate interventions.

Student Achievement. We then used topic prevalence at the school-year level to predict a school's current-year mean achievement in math and reading separately. The model specifications are identical to our analyses on absenteeism. We observed that Topic 15 ("setting goals for, recognizing, and monitoring teachers' and students' growth") was significantly positively correlated with increases in school-level average student achievement. A 10 percentage change increase in Topic 15 is associated with a 0.04 standard deviation increase in school average math achievement and a 0.02 standard deviation increase in school average reading achievement. This topic includes two interconnected reform strategies that prior research has found connected with student achievement gains: (a) monitoring students' progress and rewarding students based on their academic growth and (b) basing teacher incentives and dismissals on student achievement and growth.

Similar to reform activities depicted by Topic 10, when teachers monitor students' progress, teachers' decision making improves and students become more aware of their own performance, and subsequently, student achievement improves (Fuchs, Deno, & Mirkin, 1984; Safer & Fleischman, 2005). Moreover, a recent experimental study demonstrated that both financial and nonfinancial student incentives can generate substantial effects on test scores (Fryer, 2011; Levitt, List, Neckermann, & Sadoff, 2012). Besides rewarding students, as illustrated below in Topic 15, Task 3, the program that offers both monetary rewards and public recognition to teachers based on rigorous evaluations of their performance and is closely tied to student learning, has shown positive influence on teacher professional growth (e.g., Dee & Wyckoff, 2015). The reward was given in the format of advancing teachers' careers (e.g., Career Ladder program), which may have the

TABLE 5

The Associations Between Reform Topics and School Average Student Absences

Reform topics	Full-day absences	Part-day absences	% chronic full-day absences
Topic 1. Interventions and supports for promoting positive student behaviors	1.599	−3.147	0.055
	(3.177)	(3.560)	(0.082)
Topic 2. General parent and community outreach	−7.447	−2.375	−0.217
	(4.423)	(4.784)	(0.120)
Topic 3. Engaging parents about student academic and behavioral learning in schools	−3.820	−4.656	−0.090
	(2.408)	(3.233)	(0.067)
Topic 4. Planning, providing, and evaluating professional development for instructional improvement	−6.200	0.243	−0.153
	(3.361)	(5.518)	(0.096)
Topic 5. Monitoring student progress and using data to develop interventions	9.895	5.699	0.210
	(5.276)	(5.518)	(0.137)
Topic 7. Using assessment data to identify students for targeted support	3.652	1.387	0.168
	(3.971)	(4.683)	(0.120)
Topic 8. Extending instructional time and aligning curriculum/assessments to standards	7.178	0.229	0.164
	(4.585)	(4.402)	(0.114)
Topic 9. Teacher team (e.g., grade-level team, PLC) activities (e.g., reviewing data, planning, aligning standards, interventions)	−10.330***	−16.750***	−0.241***
	(2.737)	(4.364)	(0.068)
Topic 10. Administering common assessments and disaggregating data to differentiate interventions	−16.36**	−14.06*	−0.396**
	(5.867)	(6.961)	(0.144)
Topic 11. Leadership teams setting goals and reviewing data for school improvement	6.686	3.876	0.131
	(5.018)	(5.153)	(0.106)
Topic 12. Teacher instructional improvement via walkthroughs, observations, and feedback	1.701	6.552	0.0736
	(2.955)	(4.111)	(0.075)
Topic 15. Setting goals for and recognizing teachers' and students' growth	1.777	0.389	0.0472
	(3.524)	(4.316)	(0.085)
Topic 17. Extending learning time (or opportunities) for students and staff	1.651	10.32	0.0414
	(4.304)	(5.552)	(0.113)
Topic 18. Collecting, analyzing, and aligning student assessments	−2.080	7.939	0.038
	(6.625)	(8.064)	(0.177)
Topic 19. Improving special education	5.024	−1.479	0.137
	(4.834)	(6.434)	(0.123)
N	599	599	599

Note. Each reform topic was added to the model separately. The models control for schools' prior achievement, the number of tasks schools were implementing in a given year, and school characteristics. The standard errors are clustered at school level. PLC = professional learning communities.

**p* ≤ .05. ***p* ≤ .01. ****p* ≤ .001.

potential of promoting these teachers' instructional leadership roles in schools.

> Topic 15, Task 1: Students who achieve at the A and B levels will continue to be recognized as meeting honor roll or high honor roll, as they were last year at [school]. Students who do not show evidence of meeting the learning targets will earn the letter grade of an F. Students will receive a Pass or Fail in advisory and in some course work where individual education plans drive the students learning targets.

> Topic 15, Task 2: This memorandum outlines the financial incentives for staff documenting positive academic achievement gains in reading and/or math on HSPE, Benchmark assessments, End of Course Evaluations, or other data sources. [School District] and [District] Education Association cooperatively

> developed the new TGEM process for implementation during the 2011-12 school year. The district began the incentive model with one of the MERIT schools during the 2010-11 school year that demonstrated exceptional student growth on the measurements of student progress and grade level assessments. 41 staff members were awarded a commemorative plague and a catered luncheon on May 31, 2012. During this school year, one math teacher and one English teacher were replaced due to poor student achievement results . . .

> Topic 15, Task 3: Year 1 update: This year there is a district wide system. Those teachers rated as innovative were given the opportunity to access the career ladder. Next year we will be able to have two mentor teachers in our building . . . Year 3 update: all teachers received school-funded monetary rewards in acknowledgement of the dramatic improvements in graduation rates, state assessments in math, and end-of-course examinations in science; improvements exceeded the school's goals in these areas. Additionally, all teachers receive apparel with their academy logo as acknowledgement of their work within their academy and the progress of their students . . .

Two other topics are significantly negatively correlated with achievement in math: Topic 1 ("interventions and supports for staff and students to promote positive student behaviors"; $\beta = -.4$, $SE = 0.17$) and Topic 2 ("general parent and community outreach"; $\beta = -.52$, $SE = 0.187$). As illustrated in the exemplary tasks below, tasks in these two topics are often written in a general way rather than specifically focusing on student academic learning. The type of parent engagement activities depicted in Topic 2 may actually sidetrack schools' efforts on improving student learning and may divert school-based resources (Epstein, 1995). As we discussed further in our "Discussions and Conclusion" section, these negative associations may be due to the features of tasks analyzed in this study and may not indicate the ineffectiveness of these reform topics. The online Supplemental Table S6 includes three exemplary tasks for each topic to facilitate interpretations.

> Topic 1, Task 1: The PBIS committee will meet monthly to plan for teaching school-wide values in each classroom and celebrations for students.

> Topic 1, Task 2: Teachers will introduce and teach the 3R's (I treat others with RESPECT, I am RESPONSIBLE and I REFLECT on my choices) and model three behaviors that go with each by November 15, 2013.

> Topic 1, Task 3: Individual classroom positive enforcers: SAM tickets (good behavior is rewarded with tickets to use at lunch and school store). PAX positive classroom management system.

> Topic 2, Task 1: A community outreach dinner will be held this year to bring the community back into the schools to see what is happening and build community participation in the schools.

> Topic 2, Task 2: The Family Community Outreach Committee is actively recruiting parents to be involved in school events. Through this process, the goal is to invite parents to be a part of school improvement planning in the future years.

> Topic 2, Task 3: Monthly parent meetings provided opportunities for families to connect to the school. The school took steps to increase the amount of communication going out to parents (online, newsletters, mailings home, etc.), although the majority of communication was still one-way. The creation of a family support specialist who works with the counseling department to identify and support families and students who are struggling became a significant tool for connecting with families who had students at risk of not graduating.

Discussions and Conclusion

Using text data from underperforming Washington State schools' improvement planning and implementation reports, this article demonstrates the opportunities afforded by novel "big data" sources to study the processes of change. The LDA text analysis method we used efficiently extracted 15 school improvement strategies from the report texts that are aligned with several aspects of the policies governing school reform efforts at SIG, priority, and focus schools. The prevalence of these school improvement strategies varies greatly across schools and shows high alignment with the reform priorities self-reported by school leaders during interviews. Moreover, some of the measures are associated with increases in student achievement and reductions in student absenteeism in directions that are consistent with prior literature. For example, one cut-through theme across Topics 9, 10, and 15 that are significantly associated with either the reduction in student absences or test score gains is teachers' use of data to inform

TABLE 6

The Associations Between Reform Topics and Student Achievement

Reform topics	Math	Reading
Topic 1. Interventions and supports for promoting positive student behaviors	−0.400*	−0.150
	(0.173)	(0.122)
Topic 2. General parent and community outreach	−0.520**	−0.169
	(0.187)	(0.164)
Topic 3. Engaging parents about student academic and behavioral learning in schools	−0.015	0.0008
	(0.137)	(0.116)
Topic 4. Planning, providing, and evaluating professional development for instructional improvement	−0.282	−0.055
	(0.199)	(0.155)
Topic 5. Monitoring student progress and using data to develop interventions	−0.168	0.372
	(0.176)	(0.195)
Topic 7. Using assessment data to identify students for targeted support	−0.358	−0.261
	(0.359)	(0.147)
Topic 8. Extending instructional time and aligning curriculum/ assessments to standards	0.290	0.201
	(0.233)	(0.136)
Topic 9. Teacher team (e.g., grade-level team, PLC) activities (e.g., reviewing data, planning, aligning standards, developing interventions)	0.013	−0.357
	(0.121)	(0.232)
Topic 10. Administering common assessments and disaggregating data to differentiate interventions	0.513	0.297
	(0.269)	(0.265)
Topic 11. Leadership teams setting goals and reviewing data for school improvement	−0.096	−0.090
	(0.167)	(0.147)
Topic 12. Teacher instructional improvement via walkthroughs, observations, and feedback	−0.010	−0.218
	(0.158)	(0.135)
Topic 15. Setting goals for and recognizing teachers' and students' growth	0.439*	0.230*
	(0.201)	(0.107)
Topic 17. Extending learning time (or opportunities) for students and staff	−0.128	−0.025
	(0.307)	(0.174)
Topic 18. Collecting, analyzing, and analyzing student assessments	0.038	0.183
	(0.269)	(0.247)
Topic 19. Improving special education	0.023	−0.021
	(0.271)	(0.294)
N	596	580

Note. Each reform topic was added to the model separately. The models control for schools' prior achievement in math and reading, the number of tasks schools were performing in a given year, and school characteristics. The standard errors are clustered at school level. PLC = professional learning communities.
*$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

instruction and develop targeted interventions for at-risk students. The other cut-through theme is setting improvement targets for both students and teachers and providing incentives for meeting these targets. Reform strategies that are negatively associated with student outcomes are ones that divert school resources away from the focus of student academic learning. These findings are consistent with prior literature as we already discussed in the "Results" section; yet, the added value of our work to the literature is the detailed, specific attributes of these reform strategies that text data reveal about these themes, particularly in school improvement contexts.

This study also serves as a proof of concept that detailed textual data, particularly when linked to conventional administrative data about program outcomes and contexts, offer a promising opportunity for researchers to explore key processes of change. A more in-depth understanding of reform processes may then support practitioners in developing evidence-based theories of action for reforms and enacting positive changes in schools.

Although promising, however, the text-as-data approach to reform analysis requires caution on the part of researchers. As illustrated in this study, text data themselves and computer-assisted text analysis results need extensive validation. We used interview data to interrogate the credibility of the text data; we also used human coding and the relationships between our identified measures and student outcomes to demonstrate that our identified measures of school improvement strategies are likely conceptually valid. Despite these efforts, our results might be still limited by the nature of the reports and text analysis methods themselves. For example, the reports in this study were nonconsequential, and, thus, schools might spend less time and energy on providing consistent and accurate reporting. The relatively low quality of reporting may explain why some reform strategies that were significantly associated with student outcomes in prior studies (e.g., engaging parents about student academic and behavioral learning in schools; Rogers et al., 2017; Rogers & Feller, 2018) lack associations in our study. In this sense, if the reporting becomes more consequential, one can imagine that schools may provide higher quality reporting, and reports can be audited against lofty writing. In this context, text analysis might become a more useful tool in facilitating researchers and policymakers to monitor implementation. Moreover, these nonsignificant associations between reform topics and school performance might be because the topic proportions are imperfect measures of schools' priorities, or because we did not identify appropriate student outcome measures (e.g., correlating promoting positive student behaviors with discipline referrals). In other words, these nonsignificant associations do not necessarily indicate that these topics/reform strategies are ineffective; rather, our study might be limited based on the types of tasks that were used by these schools for each topic.

In addition, the varying quality of implementation can explain why some reform strategies were not associated with school performance as well. As evidenced in our interviews of school principals, many statistically insignificant topics were implemented with great variation across schools. For example, it appears a consensus among interviewed principals that "it is hard to implement PBIS [Positive Behavior Interventions and Supports, Topic 1] successfully." Moreover, we conducted one interview with a coach who was assigned by the state to support nine schools identified for improvement. He shed further light on many factors that could affect the implementation quality across his caseloads, such as the extent to which school leadership ensured the integrity of implementation, how committed school staff were to students' growth, and constraints from collective bargaining agreements in local school districts. It would be fruitful for future studies to more thoroughly examine these implementation issues.

Because the reports may not fully capture this varying implementation quality in schools, again, we warn policymakers and researchers that they need to be cautious of using text analysis and reports for consequential decisions. Given the limitations of this quantitative approach, qualitative case studies would certainly help deepen our understanding of reform implementation in ways that could further tease out the associations between reform strategies and outcomes. Overall, our demonstration of this method shows that although having great promise, automated text analysis methods require researchers to thoughtfully interrogate the data and each analytic step to make appropriate modeling decisions to avoid potential pitfalls. Importantly, this approach relies on researchers to use discipline-specific knowledge to interpret the results.

In substance, this study contributes to the very thin literature on the planning of school organizational improvement. Although there are broader debates on the importance of planning for organizational improvement in noneducational settings (Grinyer, Al-Bazzaz, & Yasai-Ardekani, 1986; Miller & Cardinal, 1994; Mintzberg, 1994; Spee & Jarzabokski, 2011), little work in education policy empirically associates school improvement planning with student performance. Fernandez's (2011) study of 303 school improvement plans suggested a strong and consistent association between plan quality and school-level student math and reading scores, and Strunk et al. (2016) found a somewhat positive association between plan quality and principals' reported intermediate outcomes, whereas Mintrop and

MacLellan (2002) found a null association between those planned activities and student performance. Our work extends these prior studies by directly linking school reform activities that are both planned and implemented to student outcomes using a more diverse school sample on a much broader scale than previous studies. Statistically significant associations between several reform strategies identified using reports and student performance in our study are in contrast with Mintrop and MacLellan's (2002) study of the null effects of planned activities on school performance. One possible explanation for this difference may be that reform activities captured in our study are not only planned but also implemented according to schools' reporting.

Finally, improving underperforming schools continues to be at the center of the ESSA of 2015, and serves as a critical policy lever for reducing educational inequality. The fact that such dramatic school transformation is also very costly makes it even more critical to learn what reform practices work and how they work. Given the limitations of the data and methods discussed previously, the identified associations between several reform strategies and student outcomes are not causal, and the measurement of nuanced school reform strategies also has a lot of room for improvement. With the rapid development of machine learning and deepened interests in applying those tools in education policy studies, we hope that future research can build on these exploratory findings to continue investigating effective reform strategies in more causally rigorous settings.

Acknowledgments

We thank Susanna Loeb, John Wilkerson, and discussants and participants of Society for Research on Educational Effectiveness and Association for Public Policy Analysis and Management annual conferences for their comments and suggestions on the earlier drafts.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported by grants from Spencer Foundation and the Royal Research Funds at University of Washington. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Notes

1. The achievement index for elementary and middle schools uses a 60% growth and 40% proficiency weighting. Growth is estimated by student growth percentile. For high schools, this measure also includes the 5-year adjusted graduation rate.

2. We contrasted the complete and incomplete tasks and results are included in the online Supplemental Table S2. To note here, reform topics in this article were estimated using only complete tasks; therefore, we do not know the nature of these incomplete tasks. We do know that schools serving higher proportions of students of color, or students from low socioeconomic status families, or academically underperforming students, had higher average rates of task completion.

3. The final analytic sample includes 25 School Improvement Grant (SIG) annual reports from 17 SIG schools; these represent 43% of the total number of 58 reports from 26 SIG schools. We have eight SIG schools that only had annual reports. Moreover, we have some school years that we have both annual reports and Comprehensive School Improvement Planning and Implementation Reports (CSIPIRs; 19 school-year observations). We compared the means of topic proportions between these two types of reports for the same school at the same year. They are highly comparable, which indicates that the contents of these types of reports are similar. As noted, in cases where we have both types of reports, we prioritize the CSIPIRs to increase consistency of data sources. Although SIG annual reports are not perfectly identical with CSIPIRs, they provide an alternative data source to replace missingness of CSIPIRs. Addressing the missingness improves the precision of estimation, particularly for SIG schools.

4. Unfortunately, our data sets do not have information on days of student presences or tardiness, which prevents us from calculating the total number of school days in a student's year and, thus, the percentage of the school year missed. Despite this concern, our measures might still provide a good proxy of central tendency of student absenteeism at the school level. Chronic absenteeism is commonly defined as missing 10% or more of a school year. However, the U.S. Department of Education (USED) used the proportion of students who were absent 15 or more days of the school year when reporting chronic absenteeism in the 2013–2014 Civil Rights Data Collection (CRDC).

5. One coder is a graduate research assistant who used to be a classroom teacher and has extensive experience working with school principals on planning and implementing school improvement efforts across the state. He is also the person who conducted the principal interviews to validate the text data. The other coder is a faculty researcher who has been collaborating with districts and schools to codesign school improvement plans using large-scale administrative data. She has published on the topic of school improvement and has a thorough understanding of the related literature.

6. The topic numbering is random and has no particular meaning. This order represents the unsupervised nature of the text analysis modeling, whereas our topic labels and topic coherence ratings represent the supervised elements—namely, researchers' interactions with the data and sense making of the measures.

7. Although the coefficient estimates of this measure (about 0.0003) are small and the estimates of other variables in the models are not much influenced by adding this variable, this measure is conceptually sound.

8. We controlled for prior achievement to maximize our sample size because Washington State started to collect attendance data from the 2012–13 school year.

9. As our analysis is not causal, our estimates likely suffer from omitted variable bias. The direction of the bias depends on the relationship between the omitted variables and the focal topic. In addition, all of the topics we estimated sum up to 1 at the school-year level, meaning that schools have limited resources and the increase of investment in one initiative means the decrease of investment in other initiatives. The correlational nature of our coefficients constrains our ability to suggest ways in which schools can maximize the overall reform effect by using a combination of strategies.

References

Allensworth, E. M., & Easton, J. Q. (2005). *The on-track indicator as a predictor of high school graduation*. Consortium on Chicago School Research. Retrieved from https://consortium.uchicago.edu/sites/default/files/2018-10/p78.pdf

Aulck, L., Malters, J., Lee, C., Mancinelli, G., Sun, M., & West, J. (2019, May). *Helping students FIG-ure it out: A computational mixed-methods study of freshmen seminars via FIGs*. Paper presented at the annual meeting of Society for Research on Educational Effectiveness (SREE), Washington, DC.

Balfanz, R., & Byrnes, V. (2013). *Meeting the challenge of combating chronic absenteeism: Impact of the NYC mayor's interagency task force on chronic absenteeism and school attendance and its implications for other cities*. Baltimore, MD: Johns Hopkins School of Education.

Beattie, G., Laliberté, J. W. P., & Oreopoulos, P. (2018). Thrivers and divers: Using non-academic measures to predict college success and failure. *Economics of Education Review*, *62*, 170–182.

Bettinger, E., Liu, J., & Loeb, S. (2016). Connections matter: How interactive peers affect students in online college courses. *Journal of Policy Analysis and Management*, *35*, 932–954.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bonilla, S., & Dee, T. (2017). *The effects of school reform under NCLB waivers: Evidence from focus schools in Kentucky* (No. w23462). Cambridge, MA: National Bureau of Economic Research.

Borman, T. H., Bos, J. M., O'Brien, B. C., Park, S. J., & Liu, F. (2017). *I3 BARR validation study: Impact findings cohorts 1 and 2*. American Institutes for Research. Retrieved from https://www.air.org/sites/default/files/downloads/report/BARR-report-cohorts-1-and-2-January-2017.pdf

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

Chan, E. K. H. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. In H. L. Mcbride, R. M. Wiens, M. McDonald, & E. K. H. Chan (Eds.), *The Edinburgh Postnatal Depression Scale (EPDS): A review of the reported validity evidence* (pp. 9–24). New York, NY: Springer.

Dee, T. (2012). *School turnarounds: Evidence from the 2009 stimulus* (No. w17990). Cambridge, MA: National Bureau of Economic Research.

Dee, T., & Dizon-Ross, E. (2017). *School performance, accountability and waiver reforms: Evidence from Louisiana* (No. w23463). Cambridge, MA: National Bureau of Economic Research.

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

Dragoset, L., Thomas, J., Herrmann, M., Deke, J., James-Burdumy, S., Graczewski, C., . . . Giffin, J. (2017). *School Improvement Grants: Implementation and effectiveness* (NCEE 2017-4013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Epstein, J. L. (1995). School/family/community partnerships. *Phi Delta Kappan*, *76*, 701–712.

Fernandez, K. E. (2011). Evaluating school improvement plans and their effect on academic performance. *Education Policy*, *25*, 338–367.

Fryer, R. G., Jr. (2011). Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics*, *126*, 1755–1798.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal*, *21*, 449–460.

Gottfried, M. A. (2010). Evaluating the relationship between student attendance and achievement in urban elementary and middle schools: An instrumental variables approach. *American Educational Research Journal*, *47*, 434–465.

Grimmer, J., Messing, S., & Westwood, S. J. (2012). How words and money cultivate a personal vote: The effect of legislator credit claiming on constituent credit allocation. *American Political Science Review*, *106*, 703–719.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*, 267–297.

Grinyer, P. H., Al-Bazzaz, S., & Yasai-Ardekani, M. (1986). Toward a contingency theory of corporate planning: Findings in 48 UK companies. *Strategic Management Journal*, *7*, 3–28.

Guthrie, J. E., & Henry, G. T. (2016, November). *When the LATE ain't ATE: Comparing alternative methods for evaluating reform impacts in low-achieving schools*. Paper presented at the Annual meeting of the Association for Public Policy Analysis and Management, Washington, DC.

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, *11*, 1–21.

Heissel, J. A., & Ladd, H. F. (2018). School turnaround in North Carolina: A regression discontinuity analysis. *Economics of Education Review*, *62*, 302–320.

Hickman, G. P., Bartholomew, M., & Mathwig, J. (2007). *The differential development trajectories of rural high school dropouts and graduates*. Phoenix, AZ: The College of Teacher Education and Leadership at the Arizona State University at the West Campus.

Hong, G., & Nomi, T. (2012). Weighting methods for assessing policy effects mediated by peer change. *Journal of Research on Educational Effectiveness*, *5*, 261–289.

Hurlburt, S., Le Floch, K. C., Therriault, S. B., & Cole, S. (2011). *Baseline analyses of SIG applications and SIG-eligible and SIG-awarded schools* (NCEE 2011-4019). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hurlburt, S., Therriault, S. B., & Le Floch, K. C. (2012). *School Improvement Grants: Analyses of state applications and eligible and awarded schools* (NCEE 2012-4060). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*, 451–464.

Kim, I. S. (2017). Political cleavages within industry: Firm-level lobbying for trade liberalization. *American Political Science Review*, *111*, 1–20.

Lachat, M. A., & Smith, S. (2005). Practices that support data use in urban high schools. *Journal of Education for Students Placed at Risk*, *10*, 333–349.

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2012). *The behavioralist goes to school: Leveraging behavioral economics to improve educational performance* (No. w18165). Cambridge, MA: National Bureau of Economic Research.

Liu, J. (2017, November). *Looking into classrooms: Using text-as-data methods to understand beneficial teacher practices at scale*. Paper presented at the annual meeting of the Association for Public Policy Analysis and Management (APPAM), Chicago, IL.

May, H., & Supovitz, J. A. (2006). Capturing the cumulative effects of school reform: An 11-year study of the impacts of America's choice on student achievement. *Educational Evaluation and Policy Analysis*, *28*, 231–257.

Miller, C. C., & Cardinal, L. B. (1994). Strategic planning and firm performance: A synthesis of more than two decades of research. *Academy of Management Journal*, *37*, 1649–1665.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). Edinburgh, Scotland, UK: Association for Computational Linguistics.

Mintrop, H., & MacLellan, A. M. (2002). School improvement plans in elementary and middle schools on probation. *Elementary School Journal*, *102*, 275–300.

Mintzberg, H. (1994). The fall and rise of strategic planning. *Harvard Business Review*, *72*, 107–114.

Papay, J. (2015, November). *The effects of school turnaround strategies in Massachusetts*. Paper presented at the annual meeting of the Association of Public Policy and Management. Miami, FL.

Penner, E., Rochmes, J., Liu, J., Solanki, S., & Loeb, S. (2019). Differing view of equity: How prospective educators perceive their role in closing achievement gaps. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *5*, 103–127.

Raudenbush, S. W., Reardon, S. F., & Nomi, T. (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of Research on Educational Effectiveness*, *5*, 303–332.

Reardon, S. F., & Raudenbush, S. W. (2013). Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociological Methods & Research*, *42*, 143–163.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., . . . Rand, D. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*, 1064–1082.

Robinson, C., Lee, M. G. L., Dearing, E., & Rogers, T. (2018). Reducing student absenteeism in the early grades by targeting parental beliefs. *American Educational Research Journal*, *55*, 1163–1192.

Rogers, T., Duncan, T., Wolford, T., Ternovski, J., Subramanyam, S., & Reitano, A. (2017). *A randomized experiment using absenteeism information to "nudge" attendance* (REL 2017–252). Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/edlabs

Rogers, T., & Feller, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Human Behaviour*, *2*, 335–342.

Safer, N., & Fleischman, S. (2005). Research matters: How student progress monitoring improves instruction. *Educational Leadership*, *62*, 81–83.

Saldana, J. (2009). An introduction to codes and coding. In J. Saldana (Ed.), *The coding manual for qualitative researchers* (pp. 1–31). Thousand Oaks, CA: SAGE.

Sinclair, M. F., Christenson, S. L., Evelo, D. L., & Hurley, C. M. (1998). Dropout prevention for youth with disabilities: Efficacy of a sustained school engagement procedure. *Exceptional Children*, *65*, 7–21.

Sinclair, M. F., Christenson, S. L., & Thurlow, M. L. (2005). Promoting school completion of urban secondary youth with emotional or behavioral disabilities. *Exceptional Children*, *71*, 465–482.

Singer, J. D. (2018). Even more challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, *11*, 22–24.

Spee, A. P., & Jarzabokski, P. (2011). Strategic planning as communicative process. *Organization Studies*, *32*, 1217–1245.

Strunk, K. O., Marsh, J. A., Bush-Mecenas, S. C., & Duque, M. R. (2016). The best laid plans: An examination of school plan quality and implementation in a school improvement initiative. *Educational Administration Quarterly*, *52*, 259–309.

Sun, M., Penner, E., & Loeb, S. (2017). Resource- and approach-driven multi-dimensional change: Three-year effects of School Improvement Grants. *American Educational Research Journal*, *54*, 607–643.

Wang, C., Paisley, J., & Blei, D. (2011, June). Online variational inference for the hierarchical Dirichlet process. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics* (pp. 752–760). Fort Lauderdale, FL: AISTATS.

Wang, Z., Miller, K., & Cortina, K. (2013). Using the LENA in teacher training: Promoting student involvement through automated feedback. *Unterrichtswissenschaft*, *41*, 290–305.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*, 778–808.

Whitney, C. R., & Liu, J. (2017). What we're missing: A descriptive analysis of part-day absenteeism in secondary school. *AERA Open*, *3*(2). doi:10.1177/2332858417703660

Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, *20*, 529–544.

Wilkerson, J., Smith, D., & Stramp, N. (2015). Tracing the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science*, *59*, 943–956.

Zimmer, R., Henry, G. T., & Kho, A. (2017). The effects of school turnaround in Tennessee's achievement school district and innovation zones. *Educational Evaluation and Policy Analysis*, *39*, 670–696.

Authors

MIN SUN is an associate professor in education policy in the College of Education at the University of Washington. Her work uses quantitative methods to study educator quality, school accountability, and school improvement.

JING LIU is a postdoctoral research associate at the Annenberg Institute, Brown University. His current research focuses on the causes and consequences of unequal learning opportunities indicated by absenteeism and discipline infractions, the intersection of noncognitive outcomes and teacher and school quality, and the application of computational social science methods in educational policy research.

JUNMENG ZHU is a masters student in measurement and statistics in the College of Education at University of Education. Her research areas are computational social science, statistics and education policy.

ZACHARY LECLAIR is a doctoral student in educational policy at the University of Washington's College of Education. His research interests include educator labor markets, school finance, and issues of educational equity.

Manuscript received November 15, 2018
First revision received May 1, 2019
Second revision received July 1, 2019
Third revision received July 11, 2019
Accepted July 13, 2019