ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-68

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
University of Washington

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 paper 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 strategies measures were significantly associated with reductions in student chronic absenteeism and improvements in student achievement. We lastly discuss the opportunities and pitfalls of using novel text data to study reform processes

VERSION: May 2019

Suggested citation: Sun, M., Jing, L., Zhu, J., & LeClair, Z. (2019). Using a Text-as-Data Approach to Understand Reform Processes: A Deep Exploration of School Improvement Strategies (EdWorkingPaper No.19-68). Retrieved from Annenberg Institute at Brown University: http://edworkingpapers.com/ai19-68

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, and Zachary LeClair, University of Washington

Contact

Min Sun (corresponding author) 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. She can be reached via email at: misun@uw.edu, or via phone at 206-221-1625, or via fax at 206-616-6311 or via mail at: 2012 Skagit Lane, M205 Miller Hall (Box 353600), Seattle, WA 98195.

Acknowledgement

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. In addition, we wish to thank Susanna Loeb, John Wilkerson, and discussants and participants of SREE and APPAM annual conferences for their comments and suggestions on the earlier drafts.

Abstract

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 paper 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 strategies measures were significantly associated with reductions in student chronic absenteeism and improvements in student achievement. We lastly 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 educational 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 educational 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 paper 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 datasets 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 paper 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, however, offer promising solutions to such 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 supervised classification methods to analyze over 170,000 House press releases and examine legislators' credit-claiming behavior, wherein legislators associate themselves with

spending in their constituent districts in order 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 paper, 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 coherent 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 sound. Using interview data, we also found that the identified reform strategies were largely consistent with school leaders' own perceptions of reform priorities. Lastly, 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. Lastly, 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 Educational 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 educational 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 et al. (2016) 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 in these nonacademic measures. In a similar vein, Liu (2019) used a structural topic model to code teachers' values and beliefs about student achievement gaps by using essays written by over 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 micro-classroom 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.(2018) 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 over 12,000 first-year interest group students' open-ended survey responses, the author 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 micro-processes of teaching, meanwhile, Kelly et al. (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. While each of these two studies focused on only one dimension of teaching, Liu (2019) 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 educational research. In the current paper, 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 importantly, we improve on prior studies that did not as thoroughly validate the measures derived from text analysis. Since

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 three-year average for subgroups on state assessments in English language arts and math (combined) or on an adjusted five-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 three-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 five-year cohort graduation rate for all students that was less than 60%. SIG schools has 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 three years. Schools would receive supplementary funding on top of their regular budgets; SIG schools were primarily funded through federal grants, while priority and focus schools were primarily funded through state funds. SIG schools were also required to follow federally prescribed school reform models. Almost all of 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. Since 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, while Zimmer, Henry, and Kho (2015) 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 three years, Louisiana's focus school reforms had no measurable impact on school performance (Dee & Dizon-Ross, 2017), while 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 in order 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 the school reform process 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 the current level of practice and 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.

[Insert Figure 1 Here]

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 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. Further, 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. Lastly, 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 paper, 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 2011–12 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 2011–12 to 2015–16), 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.* On the other hand, 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., 2011–12). 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 2015–16 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 under seven tokens were removed (about 5.8% of the total sample), however, because they provide little useful information on what the 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 this data.

Text Analysis

Schools conceived 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 *latent Dirichlet allocation* (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 $\pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK})$. 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") 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 description level. We only used unique entries so that tasks carried over from one year to the next would not distort the topic modeling process.

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 Appendix Table A4 and A5). The results are largely similar to the findings using the weighted measures in Table 4 and 5. However, the coefficients of weighted measures in Table 4 and 5 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:

$$\boldsymbol{p}_c = \sum_{k=1}^{k=n} \mathcal{p}_{k,c} * \mathcal{w}_k,$$

where $\boldsymbol{p}_c$ is the proportion for topic c at the school-year level, $\mathcal{p}_{k,c}$ is the proportion of task k on this topic, and $\mathcal{w}_k$ 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, $\mathcal{w}_k$ is calculated using the number of words in all the tasks articulated in the first two years of reform. Thus, $\boldsymbol{p}_c$ is calculated as the cumulative task proportion on topic c in the first two 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 topic 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 AERA/NCME/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 datasets 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 or parents 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].

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 3–8 and 11, Washington State's Measurements of Student Progress tests, and end-of-course exams in grades 9–12, 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 2010-11 through 2015-16. In 2010–11, 18 SIG schools were identified; however, because the state did not adopt the Indistar system until 2011–12, we do not have their CSIPIRs or SIG annual reports. Although in 2011–12 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 2012–13, 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 Table A1 of Appendix A, the final analytic sample is representative of all identified schools in terms of pre-reform characteristics and performance.

[Table 1 Here]

Table 2 summarizes the characteristics and performance of these identified schools as well as non-reform 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.

[Table 2 Here]

Results

Model Diagnostics Statistics

The LDA approach requires researchers to specify the number of topics. OSPI specified seven principles of school improvement; Bryk and colleagues (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 10–30 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 2(a), the *y*-axis indicates log probabilities. Large negative values indicate that top words do not co-occur often, while 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, while the 15- and 20- topic models are slightly worse and the 25- and 30- topic models substantially worse.

[Figure 2 here]

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 2(b), 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 models. 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. To assess the content validity of our topics, we recruited two experts with both deep knowledge of the prior literature and practical experience in K–12 schools. 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 Appendix Table A3; Mimno, Wallach, Talley, Leenders, & McCallum, 2011). Using a scale of 1–4, the two experts coded the 15-, 20-, and 30-topic models. Interrater reliability, as measured by Krippendorff's *alpha*, ranged from 0.81 to 0.89, depending on the individual topic model.

After generating individual coherence ratings for each topic in each model specification, the two coders discussed their ratings. Most of the differences were only a 1-point difference. (For example, one coder rated a topic a 2, while 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.)

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 has a rating of 3 or higher, while 73% of ratings in the 15-topic model are rated 3 or higher. More significantly, 45% of topics in the 20-topic model has a rating of 4, compared with only 13% of topics in the 15-topic model. Since an average coherence rating lower than 3 for a given topic casts doubt on 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.

[Table 3 Here]

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. 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") ($\mu = 0.70$, $SD = 0.068$); Topic 3 ("engaging parents in student academic and behavioral learning in school") ($\mu = 0.084$, $SD = 0.078$); and Topic 12 ("teacher instructional improvement via walkthroughs, observations, and feedback")($\mu = 0.075$, $SD = 0.081$). Topics with low mean proportions include Topic 10 ("administering common assessments and disaggregating data to differentiate interventions") ($\mu = 0.021$, $SD = 0.039$) and

most of the topics with low coherence ratings. Moreover, we observe large variations for many topics, with standard deviations equal to two 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, while 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.

[Figure 3 here]

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 shown in Figure 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 = -0.006) because these two reform strategies do not necessarily 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 (ρ = -0.246, $p < .001$). Although both Topic 12 and Topic 15 describe strategies targeting teachers—which explains why they are correlated—schools that spent more resources and times 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, while 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"; $\rho = 0.162$, $p < .001$), 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.

[Figure 4 Here]

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 3–6 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 four of the 10 schools, the principals and staff referenced nine or 10 of the top 10 topics in the reports, and in the other six 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 1-4 years previous to the interview. 20/23 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 strengthens our confidence in the text analysis results, 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 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 part-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 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[5]. Our models also control for pre-reform achievement level[6] and school characteristics (e.g., percentages of students eligible for free- or reduced-price lunch, 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). Since 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 points 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 points 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. 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 et al., 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 to 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, 2009). 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"), depict 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, 2014). Particularly, a recent multi-site 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 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 Jeffrey High School (pseudo school name). 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. Jeffrey School District and Jefferey 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"; $\beta = -0.4$, $SE = 0.17$) and Topic 2 ("general parent and community outreach"; $\beta = -0.52$, $SE = 0.187$). 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 discussion 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. Online Appendix A6 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.

[Table 5 Here]

Discussions and Conclusion

Using text data from underperforming Washington State schools' improvement planning and implementation reports, this paper demonstrates the opportunities afforded by novel "big data" sources to study the process 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. This form of detailed data, particularly when linked to conventional administrative data about program outcomes and contexts, offers 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.

While 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, Duncan, Wolford, Ternovski et al., 2017; Rogers & Fuller, 2018), lack associations in our study. In this sense, if the reporting becomes more consequential, one can imagine that schools may provide better quality reporting, and thus text analysis might become a more useful tool in facilitating researchers and policymakers to monitor implementation. On the flip side, consequential policy settings would require that reports be audited against lofty writing. Moreover, these nonsignificant associations between reform topics and school performance might also 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 non-significant 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 non-significant 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 spoke of potential reasons for variation across his caseloads:

> "Where the strong correlation [between Indistar reports and school outcomes] is, if the principal and leadership team are truly involved in the plan there is a high correlation but where the leadership team is not as, well let me put it this way – in some districts, the CBA really inhibits school improvement because they are looking at protecting teachers rights especially when it comes to assessment and accountability, you know not holding the teacher accountable for student growth to the point where you cannot give them a bad evaluation."

> "The leadership team in one school that I am thinking of, their entire focus in on low-level management things, schedules, far more teacher needs rather than student needs. You can have a leadership team with people assigned to tasks and it is just superfluous and you can have another leadership team where people are assigned to tasks and it has a high level of integrity. It would really depend school to school."

Since the reports may not fully capture this varying implementation in schools, again, we warn policymakers and researchers that they need to be very cautious of using text analysis and reports for consequential decisions. Recognizing the limitations of this quantitative approach,

qualitative case studies would certainly help deepening 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 automated text analysis methods require researchers to thoughtfully interrogate the data and each analytic step to make appropriate modeling decisions. Importantly, this approach relies on researchers to use discipline-specific knowledge to interpret the results.

Lastly, this study contributes to the very thin literature in the planning and implementation of school organizational improvement. Although there are broader debates on the importance of planning for organizational improvement in non-educational settings (Grinyer et al., 1986; Miller & Cardinal, 1994; Spee & Jarzabokski, 2011; Mintzberg, 1994), the literature that empirically associates school improvement planning with student performance is minimal. 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, while Mintrop and MacLellan (2002) found a null association between that planned activities and student performance. Our work extends prior studies in twofold. First, our study includes a more diverse school sample on a much broader scale. Second, compared to the plans analyzed in prior studies, the CSIPIRs in our study capture not only the planning of school improvement, but also implementation to some extent. Our text analysis approach intends to capture "what" the schools actually did, rather than merely the writing quality of their improvement plan, or "what" they planned to do. The associations between several reform strategies and student performance in our study suggest the importance of planning and monitoring of implementation in promoting positive school improvement outcomes. To make

reform reporting a more effective policy tool, reporting should serve both a planning tool for outlining actions and an implementation tool for coordinating resources and monitoring reform activities.

Endnotes

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 five-year adjusted graduation rate.

2. We contrasted the complete and incomplete tasks and results are included in online appendix Table A2. To note here, reform topics in this paper 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 socio-economic status families, or academically underperforming students, had higher average rates of task completion.

3. The final analytic sample includes 25 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 CSIPIR reports (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 CSIPIR reports to increase consistency of data sources. Although SIG annual reports are not perfectly identical with CSIPIR reports, they provide an alternative data source to replace missingness of CSIPIR reports. Addressing the missingness improves the precision of estimation, particularly for SIG schools.

4. Unfortunately, our datasets 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 percent or more of a school year. However, the US 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. 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.

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

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://scoe.net/calsoap/professional_resources/Documents/on_track_indicator.pdf

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

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 (Big Data Special Section), 35*(4), 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*(Jan), 993-1022.

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

Bradshaw, C. P., Mitchell, M.M., & Leaf, P. J. (2009). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions, 1*2(3), 133-148.

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: Springer.

Dee, T. (2012). School turnarounds: Evidence from the 2009 stimulus (No. w17990). *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*(2), 267-297.

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

Dragoset, L., Thomas, J., Herrmann, M., Deke, J., James-Burdumy, S., Graczewski, C., Boyle, A., Upton, R., Tanenbaum, 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*(9), 701.

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

Freeman, J., Simonsen, B., McCoach, D.B., Sugai, G., Lombardi, A., & Horner, R. (2016). Relationship between school-wide positive behavior interventions and supports and academic, attendance, and behavior outcomes in high schools. *Journal of Positive Behavior Interventions. 18*(1), 41-51.

Fryer Jr, R. G. (2011). Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics*, *126*(4), 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*(2), 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*(2), 434-465.

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.

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*(4), 703-719.

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

Guthrie, J. E., & Henry, G. T. (2016). *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.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., Wayman, J. C., ... & Steele, J. L. (2009). Using student achievement data to support instructional decision making. United States Department of Education. Retrieved from http://repository.upenn.edu/gse_pubs/279.

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, *11*(1), 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*(3), 261-289.

Hurlburt, S., Le Floch, K.C., Therriault, S.B., and 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., and 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*(7), 451-464.

Kerr, K. A., Marsh, J.A., Ikemoto, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education, 112*(4), 496-520.

Kim, I. S. (2017). Political cleavages within industry: firm-level lobbying for trade liberalization. *American Political Science Review*, *111*(1), 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*(3), 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). National Bureau of Economic Research.

Liu, J. (2019). Measuring beneficial teacher practices at scale: A novel application of text-as-data methods.

Marsh, J.A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record, 114* (11), 0161-4681.

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*(3), 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). Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–272.

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

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*(3), 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*(2), 143-163.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., Albertson, B., & Rand, D. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4):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*(6), 1163-1192.

Rogers, T., & Feller, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. Nature Human Behaviour, 2(5), 335.

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: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs.

Safer, N., & Fleischman, S. (2005). Research matters: How student progress monitoring improves instruction. *Educational Leadership*, *62*(5), 81-83.

Saldana, J. (2009). An introduction to codes and coding. *The Coding Manual for Qualitative Researchers*, 1-31.

Sheldon, S. B., & Epstein, J. L. (2004). Getting students to school: Using family and community involvement to reduce chronic absenteeism. *School Community Journal*, *14*(2), 39.

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*(4), 465–482.

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*(1), 7–21.

Singer, J. D. (2018). Even more challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, *11*(1), 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*(2), 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(*4), 607–643.

Wang, C., Paisley, J., & Blei, D. (2011, June). Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (pp. 752-760).

Wang, Z., Miller, K., & Cortina, K. (2013). Using the LENA in teacher training: Promoting student involvement through automated feedback. *Unterrichtswissenschaft, 4*(4).

Wayman, J. C., Shaw, S., & Cho, V. (2017). Longitudinal effects of teacher use of a computer data system on student achievement. *AERA Open, 3*(1), 2332858416685534.

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*(3), 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).

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*(4), 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*(4), 670-696.

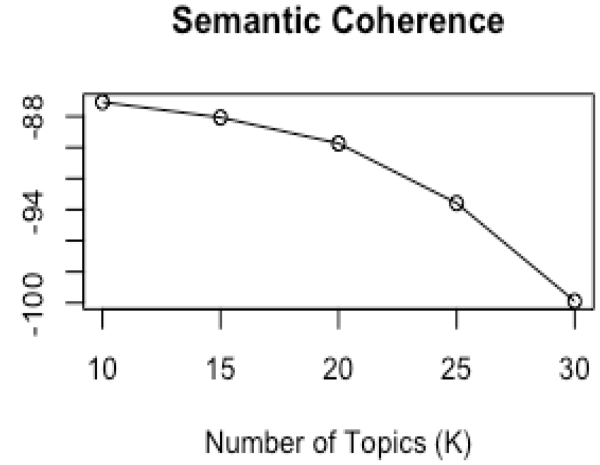
Figures and Tables

Figure 1. Examples of Tasks Written in the Comprehensive School Improvement Planning and Implementation Reports (CSIPIRs)

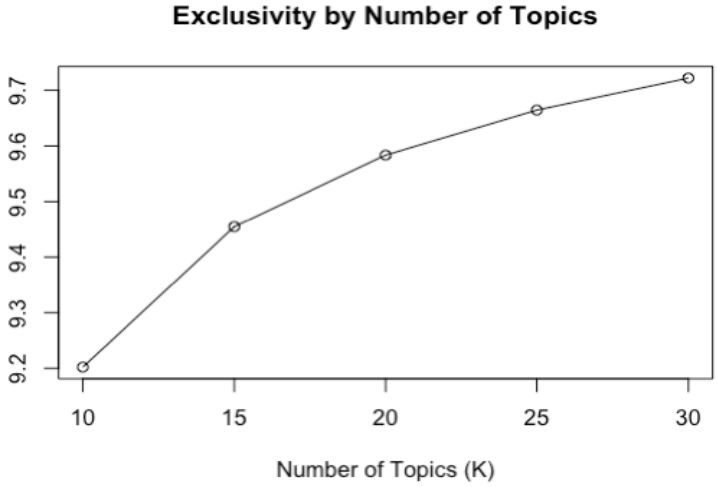
Tasks:		
1. Math team will implement learning target assessments aligned to identified grade-level power standards. Team will collaboratively review assessment data every month and adjust instruction accordingly. Team will also work to align instructional practices and math specific vocabulary during monthly meetings and weekly interventions meetings.		
	Assigned to:	Jamila Davis ← Principal
	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 ← Instructional Coach
	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

Note: Staff names are pseudonyms.

Figure 2. Topic Model Diagnostic Statistics

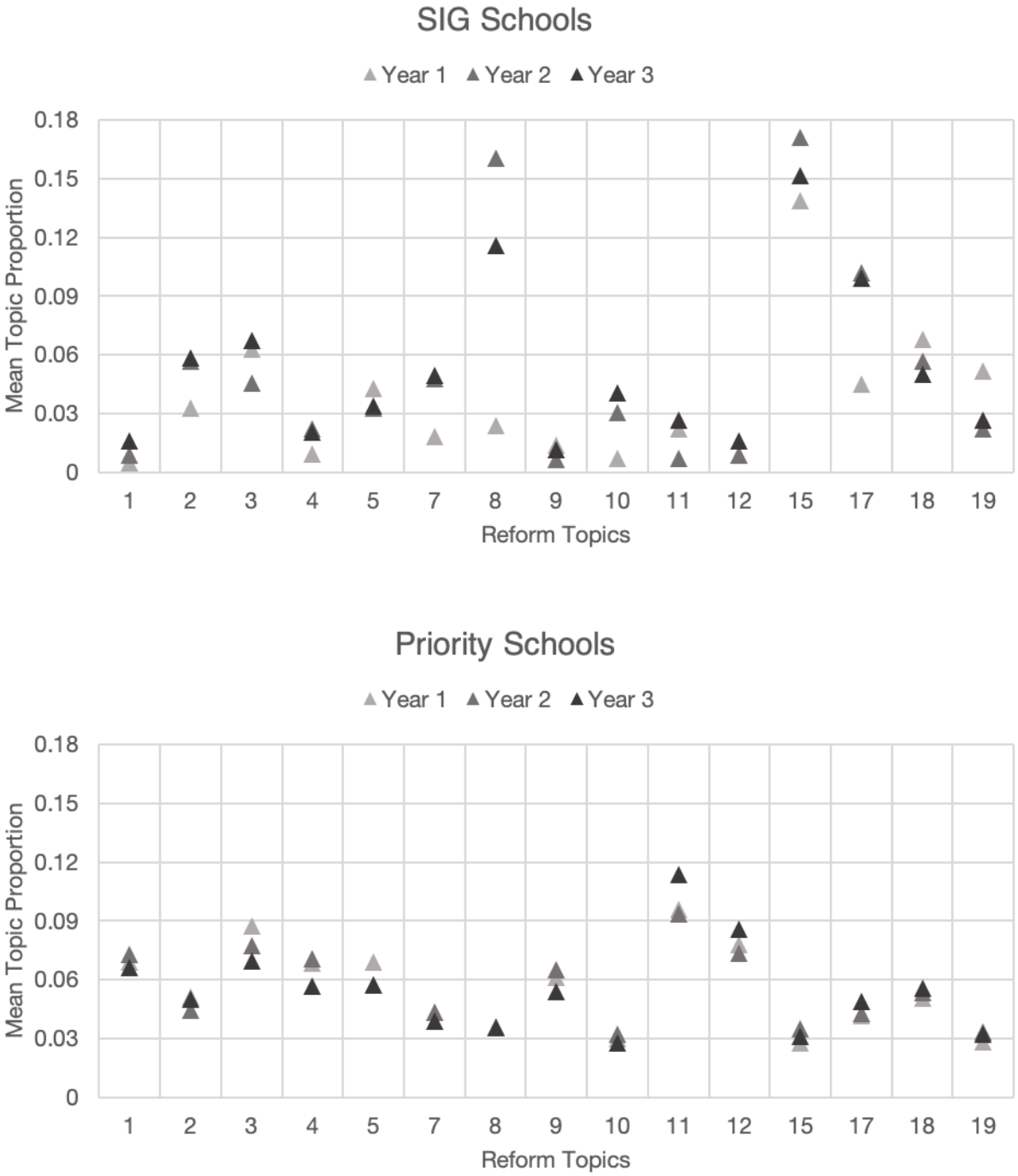


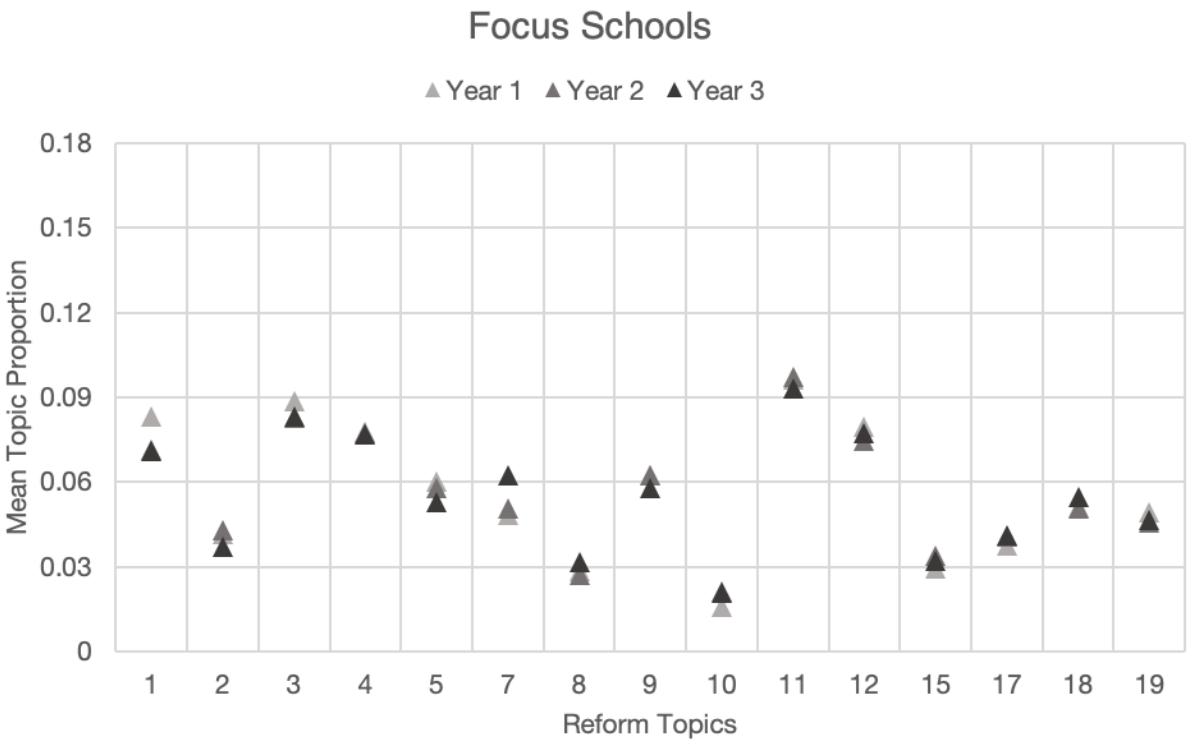
(a)



(d)

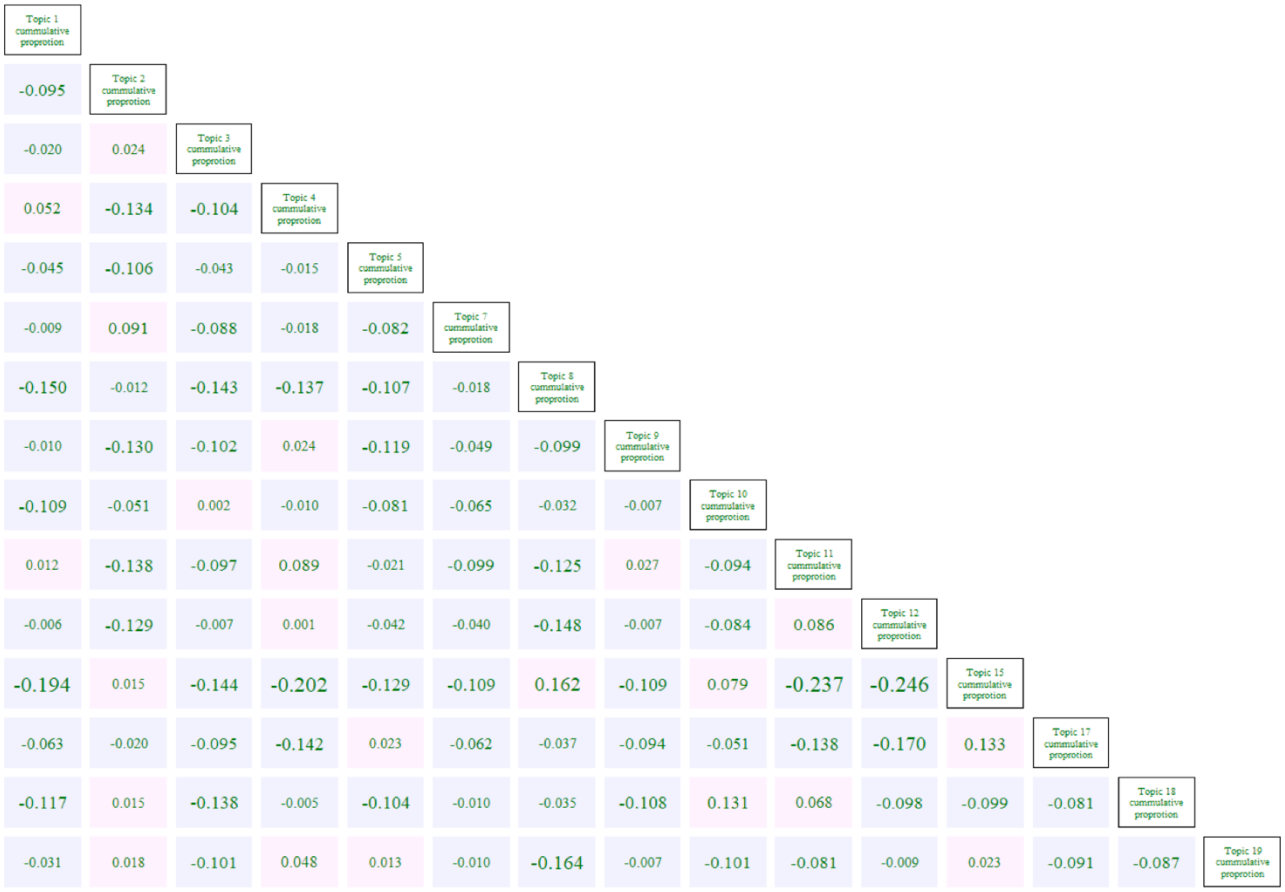
Figure 3. Topic Proportions by Reform Types and Over Reform Years





Note. The horizontal axis includes topic numbers. The corresponding labels are included in Table 3.

Figure 4. Pairwise Correlations Among Reform Strategy Topics



Note. The pink indicates positive correlation coefficients while the light purple/blue indicates negative correlation coefficients.

Table 1. Number of Treatment Schools That Have Test Score Data by Reform Type and Cohort

School Year	SIG	Priority	Focus
2015–16	0	68	116
2014–15	0	56	131
2013–14	10	34	79
2012–13	24	15	66
2011–12	24	0	0
2010–11	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). Only includes the first designation of a given school. In other words, these schools are the main analytic sample of this study.

Table 2. School-level Characteristics and Performance by Reform Type

	SIG	Priority	Focus	Non-reform
% White	0.27 (0.26)	0.39 (0.32)	0 .39 (0.25)	0.56 (0.29)
% African American	0.13 (0.19)	0.06 (0.11)	0.05 (0.09)	0 .04 (0.07)
% Hispanic	0.36 (0.32)	0.34 (0.31)	0.42 (0.26)	0.16 (0.19)
% Asian	0.06 (0.09)	0.03 (0.06)	0.04 (0.07)	0.05 (0.08)
% Pacific Islander	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)
% Native American	0.10 (0.2)	0.11 (0.22)	0.03 (0.08)	0.02 (0.08)
% multiracial	0.05 (0.05)	0.06 (0.05)	0.05 (0.04)	0.05 (0.05)
% eligible for free or reduced-price lunch	0.78 (0.19)	0.71 (0.26)	0.69 (0.17)	0.41 (0.27)
% English language learner	0.36 (0.26)	0.29 (0.28)	0.36 (0.23)	0.17 (0.19)
% homeless	0.05 (0.03)	0.06 (0.07)	0.05 (0.06)	0.03 -0.04
% special education	0.13 (0.05)	0.15 (0.11)	0.14 (0.05)	0.16 (0.21)
Prior student academic achievement	-0.6 (0.19)	-0.57 (0.30)	-0.33 (0.23)	N/A
Student academic achievement	-0.42 (0.25)	-0.47 (0.37)	-0.31 (0.29)	-0.05 (0.44)
Full-day absences	12.66 (5.85)	9.94 (6.12)	9.89 (4.67)	7.18 (5.32)
Part-day absences	5.83 (7.86)	4.09 (5.37)	4.81 (5.85)	2.89 (4.59)
% chronic absenteeism—full day	0.41 (0.16)	0.33 (0.17)	0.33 (0.12)	0.24 (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. Non-reform 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 Non-reform schools, respectively. Standard deviations are reported in parentheses.

Table 3. Descriptives of Reform Topics

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

Note. Topic proportions are at the school–year level.

Table 4. The Associations between Reform Topics and School Average Student Absences

Reform topics	Full-day absences	Part-day absences	% chronic full-day absence
Topic 1	1.599	-3.147	0.055
	(3.177)	(3.560)	(0.082)
Topic 2	-7.447	-2.375	-0.217
	(4.423)	(4.784)	(0.120)
Topic 3	-3.820	-4.656	-0.090
	(2.408)	(3.233)	(0.067)
Topic 4	-6.200	0.243	-0.153
	(3.361)	(5.518)	(0.096)
Topic 5	9.895	5.699	0.210
	(5.276)	(5.518)	(0.137)
Topic 7	3.652	1.387	0.168
	(3.971)	(4.683)	(0.120)
Topic 8	7.178	0.229	0.164
	(4.585)	(4.402)	(0.114)
Topic 9	-10.330***	-16.750***	-0.241***
	(2.737)	(4.364)	(0.068)
Topic 10	-16.36**	-14.06*	-0.396**
	(5.867)	(6.961)	(0.144)
Topic 11	6.686	3.876	0.131
	(5.018)	(5.153)	(0.106)
Topic 12	1.701	6.552	0.0736
	(2.955)	(4.111)	(0.075)
Topic 15	1.777	0.389	0.0472
	(3.524)	(4.316)	(0.085)
Topic 17	1.651	10.32	0.0414
	(4.304)	(5.552)	(0.113)
Topic 18	-2.080	7.939	0.038
	(6.625)	(8.064)	(0.177)
Topic 19	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. * $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

Table 5. The Associations between Reform Topics and Student Achievement

Reform topics	Math	Reading
Topic 1	-0.400*	-0.150
	(0.173)	(0.122)
Topic 2	-0.520**	-0.169
	(0.187)	(0.164)
Topic 3	-0.015	0.0008
	(0.137)	(0.116)
Topic 4	-0.282	-0.055
	(0.199)	(0.155)
Topic 5	-0.168	0.372
	(0.176)	(0.195)
Topic 7	-0.358	-0.261
	(0.359)	(0.147)
Topic 8	0.290	0.201
	(0.233)	(0.136)
Topic 9	0.013	-0.357
	(0.121)	(0.232)
Topic 10	0.513	0.297
	(0.269)	(0.265)
Topic 11	-0.096	-0.090
	(0.167)	(0.147)
Topic 12	-0.010	-0.218
	(0.158)	(0.135)
Topic 15	0.439*	0.230*
	(0.201)	(0.107)
Topic 17	-0.128	-0.025
	(0.307)	(0.174)
Topic 18	0.038	0.183
	(0.269)	(0.247)
Topic 19	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.

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

Online Appendix. Supplemental Tables

Table A1. Pre-reform school-level characteristics between all identified schools and the analytical sample

	All identified schools	Analytical sample
% White	0.40	0.39
	(0.27)	(0.27)
% African American	0.07	0.06
	(0.12)	(0.12)
% Hispanic	0.37	0.39
	(0.28)	(0.28)
% Asian	0.05	0.04
	(0.07)	(0.07)
% Pacific Islander	0.01	0.01
	(0.02)	(0.02)
% Native American	0.06	0.06
	(0.15)	(0.15)
% multiracial	0.04	0.04
	(0.03)	(0.03)
% eligible for free or reduced-price lunch	0.71	0.73
	(0.19)	(0.18)
% English language learner	0.33	0.34
	(0.25)	(0.25)
% homeless	0.04	0.05
	(0.04)	(0.05)
% special education	0.14	0.14
	(0.07)	(0.07)
Prior student academic achievement	-0.42	-0.43
	(0.28)	(0.27)
Full-day absences	9.32	9.63
	(5.00)	(5.12)
Part-day absences	3.62	3.80
	(4.71)	(4.94)
% chronic absenteeism—full day	0.31	0.32
	(0.13)	(0.13)
N (school–year)	772	623

Note. This table shows that the final analytic sample is representative of all identified schools in terms of pre-reform characteristics and performance. All identified schools only include schools' first reform identifications. Analytical samples are the schools with either or both outcome measures. The sample sizes reported here only reflect the analytic sample that provides all the demographic information. The sample sizes for the absence measures are 344 and 299, respectively. Standard deviations are reported in parentheses.

Table A2. Compare characteristics of schools between complete and incomplete tasks

School characteristics	Complete	Incomplete
% White*	36.889 (25.847)	39.682 (26.789)
% African American*	6.565 (11.322)	5.547 (10.044)
% Hispanic*	40.655 (28.509)	38.675 (28.014)
% Asian*	4.625 (7.577)	4.181 (7.169)
% Pacific Islander*	1.155 (1.800)	1.061 (1.751)
% Native American*	5.310 (14.398)	5.989 (14.922)
% multiracial	4.332 (3.294)	4.352 (3.265)
% eligible for free or reduced-price lunch*	72.767 (17.247)	71.232 (18.511)
% special education	13.983 (5.785)	13.891 (7.115)
Prior student academic achievement*	-0.428 (0.269)	-0.412 (0.281)
Enrollment*	501.664 (294.107)	485.395 (283.808)
Total 43,819 tasks	55.19%	44.81%

Note. * indicates statistically significant at the 0.05 level.

Table A3. Rubrics and procedures for rating topic coherence

Step 1: Preparation. Using the STM visualization graph, select at least 20 tasks that have the highest proportion loaded on one given topic. Start with the highest proportional ones, then read through the statement of tasks, synthesize key ideas across the tasks, and label each topic. Only use tasks that have > 7 words or complete sentences (or a coherent idea) to inform your labeling.

Step 2: Use the last column in the attached template in the spreadsheet to record your summary (or 2–3 key words) of the exemplary task you are reading. This helps you to clearly apply the rubrics below and keep track of tasks you have reviewed.

Step 3: Rate topic coherence using the following metrics for the extent to which the topic is coherent.

On a 4-point scale: 4 = a great deal; 3 = moderate; 2 = little; 1 = none

4 = *a great deal*: It is easy for me to find one coherent latent construct for this topic. The label emerges from the exemplary texts coherently.

3 = *moderate*: The topic contains 2–3 latent constructs; however, they are closely related. I am still able to come up with one label to summarize almost all exemplary tasks.

2 = *little*: The topic contains more than 2 latent constructs that are somewhat connected. I manage to come up a label, but it only summarizes a portion of the exemplary tasks well.

1 = *none*: The tasks under this topic is largely random, with no clear relationships.

Step 4: Record your rationales for (a) the label you have created; and (b) the topic coherence rating you have given.

Table A4. The associations between unweighted reform topics and student absences

Reform topics	Full-day absences	Part-day absences	% chronic full-day absence
Topic 1	0.335	-5.121	0.049
	(3.012)	(3.837)	(0.087)
Topic 2	-7.607	0.913	-0.290
	(5.873)	(6.184)	(0.152)
Topic 3	-3.318	-4.514	-0.092
	(2.620)	(3.873)	(0.076)
Topic 4	-2.592	3.425	-0.068
	(4.132)	(6.408)	(0.106)
Topic 5	18.140*	7.539	0.399*
	(7.473)	(6.755)	(0.180)
Topic 7	3.544	6.581	0.208
	(6.087)	(7.422)	(0.169)
Topic 8	7.475	0.026	0.165
	(5.109)	(5.232)	(0.090)
Topic 9	-13.710***	-21.980***	-0.317***
	(3.740)	(6.349)	(0.090)
Topic 10	-11.080	-17.020*	-0.310
	(9.491)	(8.409)	(0.201)
Topic 11	5.363	-0.576	0.122
	(3.852)	(5.816)	(0.099)
Topic 12	3.717	6.663	0.128
	(3.120)	(4.238)	(0.081)
Topic 15	5.334	6.877	0.150
	(6.050)	(7.686)	(0.166)
Topic 17	-0.0802	17.630*	-0.048
	(5.767)	(7.086)	(0.145)
Topic 18	-8.024	-0.252	-0.116
	(6.906)	(7.573)	(0.188)
Topic 19	5.960	-0.521	0.178
	(5.120)	(7.818)	(0.144)
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 the schools were implementing in a given year, and school characteristics. Standard errors are clustered at school level. * $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

Table A5. The associations between unweighted reform topics and student achievement

Reform topics	Math	Reading
Topic 1	-0.460*	-0.150
	(0.202)	(0.135)
Topic 2	-0.551*	0.0172
	(0.261)	(0.232)
Topic 3	-0.007	-0.039
	(0.136)	(0.111)
Topic 4	-0.164	0.102
	(0.221)	(0.172)
Topic 5	-0.208	0.524*
	(0.210)	(0.251)
Topic 7	-0.246	-0.245
	(0.401)	(0.259)
Topic 8	0.191	0.128
	(0.320)	(0.142)
Topic 9	0.046	-0.507
	(0.148)	(0.271)
Topic 10	0.764*	0.485
	(0.311)	(0.263)
Topic 11	-0.083	-0.076
	(0.194)	(0.158)
Topic 12	-0.011	-0.156
	(0.173)	(0.147)
Topic 15	0.618*	0.387*
	(0.299)	(0.193)
Topic 17	-0.048	0.007
	(0.365)	(0.207)
Topic 18	0.168	0.052
	(0.311)	(0.244)
Topic 19	0.026	0.084
	(0.285)	(0.337)
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.
* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

Table A6. Exemplary tasks for reform topics

Reform Topics	Exemplary Task 1	Exemplary Task 2	Exemplary Task 3
1. Interventions and supports for promoting positive student behaviors	[School] initiated PBIS work in August 2014. The PBIS committee will meet monthly to plan for teaching school-wide values in each classroom and celebrations for students.	Discipline Committee will look at the school wide discipline policy and incorporate strategies from the Compassionate Schools training to develop a new or revised school wide discipline plan.	Rules and procedures are modeled in all areas of the school by staff members: hallways, lunchroom, playground, parking lot, specialist classes, and classrooms
2. General parent and community outreach	In order to welcome and prepare incoming 6th graders and their parents to [school], we will have a Welcome Night held every spring.	We will implement a program we are calling "Life at [School]: Parent University." These classes will include ESL, home economics, and school success strategies. These classes will have three sessions, and each session will meet Monday through Thursday from 6pm to 8pm and last for 6 weeks.	Dual language staff will provide appointment opportunities for incoming 6th and 7th grade students and parents to visit [school] for an orientation and tour during visiting hours. The visiting parents and student will have the opportunity to see the facility as it functions during a school day.
3. Engaging parents about student academic and behavioral learning in schools	We increased parent involvement in the school environment through the creation of the Parents as Educational Partners (PEP) group. During those meetings, parents brought up the idea of uniforms, provided input on the new uniform policy, shared feedback on the schools' communication, learned about new curriculum, and got updates on student performance.	All teachers in 2014-15 will create a home to school communication plan to include student goals, strategies for home and access for parents through electronic blogging, email, or texting to support the goals set.	The principal has held parent coffee hours every other month this school year… Coffee hours are informal opportunities for families to share concerns and suggestions for school improvement… Additional opportunities for parents to be involved include student-led conferences and monthly parent nights.
4. Planning, providing, and evaluating professional development for instructional improvement	Plan at least 3 days during the year to focus on developing rigorous Performance Tasks and aligned rubrics accompanied with GLAD units that differentiate the material by ELP standards and other learning needs. Focus PLC time on this objective as well.	AVID professional development for teachers to understand and develop skills for teaching the literacy strategies identified in task 1 (KWL Charts, New Vocabulary Acquisition, and Writing in the Margins: Questioning).	Staff will complete a lesson plan form for peer-observations, making modifications to the lesson based on the team debrief conversation… These lesson plans will be submitted with a short reflection to determine whether or not instructional practices

			were changed to follow trainings.
5. Monitoring student progress and using data to develop interventions	The data specialist and principal will review the files of all entering students to look at academic and credit history to ensure a seamless transition from school to school. Supports that were implemented at the school of origin will be noted, evaluated, and when possible replicated to ease the stress of this transition.	Present last year's data and approach to truancy board. Ask for a commitment to help interpret data and guide our work in developing a non-punitive system that supports student attendance as well as providing mentors for students who need intensive interventions.	Check list/form developed and shared with all teachers for monitoring the implementation of the agreed on oral language strategies to support the development of oral language skills of identified (strategic & intensive) ELL learners.
7. Using assessment data to identify students for targeted support	During the fall quarter 4 teachers will offer additional classes after school on an extended learning schedule. These classes will offer students opportunities to take additional classes in areas where achievement data has indicated a need for additional support. These classes are: …	Developed lessons that are targeted to whole group and small group (with a focus on meeting the needs of our SWD population) after analyzing data to determine which students need what skills and what students need additional supports.	At the opening meeting prior to the 2012-13 school year, principal will share data from [tests and surveys]. Emphasis placed on subgroups performance including ELL and SPED students. Information to create initial instructional and student placement plans, as well as initial plans for interventions.
8. Extending instructional time and aligning curriculum or assessments to standards	The bell schedules (see File Cabinet, 8.1 Baseline and MERIT) shows the increase of time for classes and student instruction. The instructional calendars (see File Cabonet, 8.2 Baseline, Year One, and Year Two) show additional days added to the school year and also indicate early release days.	Throughout this year our math and literacy coaches have worked with teachers to align their curriculum materials to the state standards. From this work, pacing guides were established so that all students were given the opportunity to learn all the standards for that class or grade level.	Currently the staff is working k-12 to align vocabulary for language arts and math to Washington State Learning Standard vocabulary. The reading street curriculum has a curriculum pacing guide that is used k-5 and aligned to the Washington State Learning Standards.
9. Teacher team (e.g., grade-level team, PLC) activities (e.g., reviewing data, planning, aligning standards, developing interventions)	Grade level teams will meet every six weeks. At these meetings student data based on assessment of standards will be analyzed. Trends will be noted and revisions made to the curriculum. Items such as pacing, differentiated learning, flexible grouping, and reteaching will be implemented based on the trend data.	Principal will meet with third and fourth grade teams by grade level. Teams will review IRLA data student by student to adjust remedial groups. Teachers will review elements of SBA preparation and develop a grade level plan.	All grade levels create SMART goals for mathematics to align with the MBA. 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 been overlapping SMART goals to ensure all

			students are making progress.
10. Administering common assessments and disaggregating data to differentiate interventions	Literacy Data collected through Fountas and Pinnell assessments, spelling inventories, and Scholastic reports are used to organized students for small group instruction in reading essentials classes. Students are regrouped based on changes in performance.	As a building, we will need to determine what materials will be used for progress monitoring and formative assessment in all core subjects. Current data indicates that we have decreased the number of students performing at levels 1 & 2 from 122 (Fall) to 82 (Spring).	DIBELS testing in the fall will guide the selection of student intervention ELA groups. The interventionist and teachers will decide who will receive the interventions with interventionists and who will receive it in the classroom with their grade level team of teachers.
11. Leadership teams setting goals and reviewing data for school improvement	Team leaders will create and provide analysis of unit design/student learning outcomes to Focus School Leadership Team. Focus School Leadership Team will further analyze data and report findings to School Centered Decision Making team; take/make recommendations for further actions.	Principal will maintain a yearly calendar of SIP team meetings, PLC, meetings, Academy meetings, RI meeting and PAC meetings. In addition, vice principal, coach, counselor, and PLC leaders will also help support these meetings by creating/sending agendas for meetings and following-up with complete meeting minutes.	The leadership team will meet monthly to review [school's] SIP plan and our building's progress towards the goals outlined within our Indistar plan as documented by agendas/minutes documented in Indistar.
12. Teacher instructional improvement via walkthroughs, observations, and feedback	There will be a focus on one specific dimension of the 5 Dimensions of Teaching and Learning each month. Walkthroughs of staff classrooms will include feedback on the implementation of that specific dimension. The weekly principal's email will also focus on the specific dimension.	Create forms for learning walks and formal observations that can be integrated across platforms using Evernote. Create Evernote templates to document evidence of effective practice by individual staff members across time. Create templates to document communication of next steps/follow ups for individual staff members.	The principal will implement a walkthrough process with the staff, using a protocol focused on student engagement. He will collect data on student engagement practices and provide monthly feedback to the staff. Each month he will share a new focus for the upcoming month.
15. Setting goals for and recognizing teachers' and students' growth	Each teacher whose student growth impact rating is "high" for criterion 6 (7 or 8 score) will receive $1000 for the classroom. Recognition will also be given in the following ways: (1) Letter of recognition for their file (2) Be given public recognition	Implement a common grading scale for the entire building. B's or Better are the Target! Parents will be able to continue to access students' grades, much like in years past, through Skyward. A uniform grading scale is being implemented at [school]	During year 2 all [district] teachers wrote student growth goals as part of their evaluation. Teachers received training in support in the goal writing process, and each had to have goals approved by the principal… If teachers achieved all 3 of their

	at a school board meeting, staff meeting, and in district and school newsletters.	and many teachers will refer to it in their course syllabus this school year.	goals, they are rewarded with the attendance at a national conference of their choice.
17. Extending learning time (or opportunities) for students and staff	We added 30 additional minutes each day in order to support the extended learning time for each core course… We added additional early release days in order to accommodate increased time for teacher collaboration.	[School] will offer summer learning opportunities for students in STEM, Summer Math, and Literacy. The Literacy Program will run as a "camp" focusing on keeping students from declining in lexile level over the summer. Current students whose lexile levels have remained stagnant throughout the year will be chosen to participate.	Teachers were provided with 90 minutes per week of collaboration time through a weekly Monday late start throughout the school year. Teachers were also provided with 4 days of additional professional development time during August, and 14 hours of professional development time after school hours during the year.
18. Collecting, analyzing, and aligning student assessments	Teachers continue to utilize professional learning community time to review student data and to reflect on teaching practices. Staff members reported they are intentional with progress monitoring to inform small group interventions and use results to create settings that target student needs.	Kathy and Sandy will correlate the MBA and end of unit assessments for triangulation to present and decipher with teachers in order to plan for differentiation as well as intervention groups. Data will also be compared to classroom based data collected, i.e. exit tasks and teacher observation.	Three or more Interim Block assessments will be given in ELA and Math. Each teacher will analyze the data to adjust instruction if needed.
19. Improving special education	8th grade Gen Ed ELA teachers will meet with Special Education/Support teacher. 7th grade Gen Ed Math teachers will meet with Special Education/Support teachers. The purpose of the meeting is to bridge the communication gap between special education and gen ed.	A title 3 grant has been written (pending approval) to implement an after-school program for level 1-3 English language learners. Students will have access to Imagine Learning Experience K-2. Grades 2-6 can access ST Math, Typing, and Easy Teach at school and home. Data from these programs will be utilized to measure growth.	Psychologist, Occupational Therapist, and/or Speech and Language Pathologist will provide socio-emotional education and student classroom accommodation training to teachers, at least twice during the school year. The focus will be on inclusivity and teaching students emotional self-regulation.