Gauging Factors Associated with School Reform Task Completion: An Application of Text Analysis Methods in Policy Implementation

Junmeng Zhu

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Education

University of Washington

2019

Committee:

Min Sun

Min Li

Program Authorized to Offer Degree:

College of Education

© Copyright 2019

Junmeng Zhu

University of Washington

Abstract

Gauging Factors Associated with School Reform Task Completion: An Application of Text Analysis Methods in Policy Implementation

Junmeng Zhu

Chair of the Supervisory Committee:

Min Sun

College of Education

Prior research has suggested the importance of policy implementation, and various studies have researched how policies are put into effect (Durlak & DuPre, 2008; Schofield, 2001). However, there is limited quantitative research on understanding the factors associated with policy implementation on a large scale due to the difficulties in finding available data or collecting a large amount of data about implementation. Using the State of Washington's school improvement policy as an example, this thesis investigates what factors are associated with the completion of reform tasks across schools. This thesis, as situated in a larger project, uses information extracted from the large volume of Comprehensive School Improvement Planning and Implementation Reports (CSIPIRs) and publicly available state administrative data. Findings suggest that the content of the reform tasks and to whom the tasks were assigned are task-specific factors that are associated with completion of tasks. The results do not show a meaningful association between school contextual factors and the completion of tasks.

Introduction

Under the era of No Child Left Behind Waivers and School Improvement Grants, Washington state (WA) has applied great efforts and resources to support the underperforming schools in improving student learning. Some school improvement research studies have used various methods to investigate the mechanism by which policies or programs achieve their results (Hong & Nomi, 2012; Reardon & Raudenbush, 2013). However, these studies have their limitations in revealing the mechanism on a large scale. In the implementation literature, various factors affecting implementation have been found. Therefore, this thesis focuses on finding the factors that are associated with WA school improvement policy implementation by testing their associations with the completion of reform tasks. Whether or not reform tasks are completed is one indicator, among many others, of whether the tasks were implemented. It is important to understand the factors associated with the implementation of tasks since implementing a reform task is the first step toward achieving its outcome. Using planning reports as data and state administrative data of schools, this thesis uses text analysis methods to extract factors from reports and utilizes logistic regression to detect the factors that are associated with the completion of tasks.

Background of Washington State School Improvement

This thesis focuses on understanding the factors that are associated with school improvement implementation by evaluating the completion of reform tasks that underperforming schools planned and implemented to improve student achievement or graduation rates during the era of No Child Left Behind (NCLB). These initiatives were funded by NCLB waivers and School Improvement Grants from 2010 to 2016. Washington state supported the improvement of underperforming schools by providing grants as well as technical assistance including coaching

and monitoring progress using school written reports. During this time, WA identified underperforming schools based on their achievements in state assessments in reading and math combined, or on the Adjusted 5-year Cohort Graduation Rate for high schools. These schools received funding in addition to their regular school budgets. They also received substantial technical assistance. All the identified schools follow the state's framework of school turnaround, and they are required to submit the Comprehensive School Improvement Planning and Implementation Reports (CSIPIRs) to the State of Washington's Office of Superintendent of Public Instruction (OSPI). In each CSIPIR report, schools report their reform tasks that address different facets of reform efforts. Schools not only record their new planned tasks but also reflect the progress of implementing their ongoing tasks. Since the 2011-12 school year, all identified schools in WA have been required to submit CSIPIRs. These reports are submitted through an online platform called Indistar. When schools develop their planning reports through Indistar, they are required to list each of the specific tasks they need to implement along with information about these specific tasks, such as to whom the task is assigned, the target completion date, and if and when the task was completed. An example of reform tasks showing the frame of CSIPIR is presented in Figure 1.

[Figure 1 Here]

Purpose of Study

Although there is a significant body of research on policy and program implementation, a review of the literature revealed that it is rare to apply quantitative methods to a large dataset of reports that contain the actions of implementation and status of program implementation. Quantitative studies of factors predicting implementation differences are scarce. This thesis aims

to explore the extent to which different policy implementation factors can explain the differences in school improvement reform tasks completion.

Learning about task completion helps illuminate why some tasks are implemented while others are not and how schools can complete more tasks. All the identified underperforming schools received funds and technical assistance from the federal and/or state government. Tasks that are never completed during treatment years may waste resources and indicate a lack of rigorous and reasonable planning for the school. Thus, knowing what type of tasks can be completed during treatment time can be crucial for schools that need consistent support to achieve effective turnarounds.

Instead of finding what causes the differences, this study focuses on exploring the factors associated with task completion using measures from both the state's school administrative datasets and schools' written reports. This thesis not only contributes to implementation literature, but it also tests possible methods to extract valuable information from text data. This information can be quantified and later used in studying the factors associated with school improvement task implementation.

The key research questions can be framed in the following way:

1. What are the characteristics of the reform tasks themselves that are associated with task completion?
2. Is task completion associated with whether a task was assigned to a specific person(s)?
3. What are the school contexts, if any, that are associated with task completion?

Prior to seeking answers to these questions, this study summarizes the theoretical framework the research draws on and the factors from prior theoretical and empirical studies that affect implementations. This summary is followed by a brief description of the intended

approach to quantifying these factors. Then, in the Data and Methods section, descriptives of both the school administrative data and the textual reports data are presented. Methods for extracting information from the text data and the method for testing the effects of different factors on task completion are then introduced. Methods of topic modeling and named entity recognition (NER) used to extract information from the text data and quantify the measures are introduced. The multiple logistic regression model was used to test how the selected factors may be associated with the completion of the tasks. This section also demonstrates which features were extracted from the text data and which measures were selected for the next step of logistic regression. The selected measures are also described. The results of the multiple logistic regression are then presented, followed by a discussion of the findings and their implications.

Literature Review and Theoretical Framework

Education Reform

Although prior research studies in education policy evaluation have used various methods to probe the mechanism by which the programs achieve their effects, all of them have their limitations. A few studies have used instrumental variables and mediation analysis to study the impact of mediators on the effects of treatment. However, the assumptions for using these methods are difficult to validate, and well-designed experiments are often challenging to arrange (Hong & Nomi, 2012; Reardon & Raudenbush, 2013). Other approaches like case studies are time consuming, and data collected from interviews and observations are hard to quantify. One study used school improvement planning reports to examine the relationship between plan quality and implementation (Strunk, Marsh, Bush-Mecenas, & Duque, 2016). However, plan quality was rated by human coders. This research investigated implementation factors using text data but is unable to analyze the data at a large scale due to the time and cost constraints (Strunk

et al., 2016). To address the limitations of the previous education program implementation research, this thesis uses results generated in a larger project using topic modeling as the implementation factors. This study also applies a Natural language processing method on the text data—all the Washington state school improvement planning reports—so as to study school reform implementation on a large scale. These reports contain rich information about school reform planning and implementation. Schools use written reports to set reform goals, design reform strategies, coordinate efforts among key actors, and monitor implementation processes (Strunk et al., 2016). Using text analysis methods, this information can be quantified and related to school reform tasks implementation.

Implementation Research

The outcome of a program or policy is dependent on its level of implementation. Much of the research on program and policy implementation has confirmed that higher levels of implementation lead to better outcomes (Durlak & DuPre, 2008). However, failure in implementing a program can not only diminish its impact but also sometimes lead to negative results if the program is not implemented sufficiently (Durlak & DuPre, 2008). In the implementation literature, much research has been done to find which factors are most significant in implementation (Schofield, 2001). It is difficult to produce a positive outcome from a program without knowing the factors impacting its implementation. In addition to its impact on the outcome, knowing policy implementation factors is essential to efficient resource allocation. Research indicates that the combination of resources offered to support a program or policy is an important factor that affects implementation (Fixsen, Naoom, Blase & Friedman, 2005; Stith et al. 2006). Sufficiently implemented programs represent committed actions and full use of allocated resources. However, insufficiently implemented programs are a waste of human

and financial resources. In this context, it is worthwhile to explore the factors associated with implementation in order to allocate resources more efficiently. Moreover, understanding the process of implementation is of significance in enabling practitioners to provide support and feedback for future program implementation. Although researchers are able to use experimental and quasi-experimental designs to identify changes in outcome for a specific program or policy and report whether the program had positive impacts, this type of research often cannot uncover the mechanism by which the programs achieve their effects (Hedges, 2018). In light of the importance of understanding the factors associated with successful implementation, this thesis studies these factors by analyzing the school reports along with state administrative data of the schools in reform.

Factors Affecting Implementation

A sufficient amount of prior research has been completed on finding the factors that explain variations in program or policy implementation. These factors can be summarized into three dimensions: policies, people, and places (Honig, 2006). Past implementation research has indicated that characteristics of the policy itself affect implementation. Clear and consistent goals of a policy have great impact on its successful implementation. Success of implementation has been associated with policies that had prioritized objectives (Berman, 1978; Bullock, 1980; Ripley & Franklin, 1982; Schofield, 2001). According to Honig (2006), understanding the objectives of a policy is essential to better understand the challenges that a policy might face. Moreover, understanding the nature of a program or policy contributes to successful implementation. The nature and type of a program or policy is a variable that affects the success of implementation (Schofield, 2001). Besides the goal and nature of a policy, whether a policy aims at short- or long-term impacts also brings different challenges to implementation. Policies

targeting short-term changes have different effects on implementation than those that aim for a longer time to implement (Honig, 2006). It is not difficult to expect that resources allocation, time frame, and schedule for implementation will be different between two types of policies. Prior research has also indicated that ongoing training and technical support is a factor affecting implementation (Durlak & DuPre, 2008; Fixsen et al., 2009; Tichnor-Wagner et al., 2018).

Many research studies have also pointed to local contextual factors that affect implementation. These are characteristics of organizations in which a policy is implemented. Each organization's nature and type are also considered as variables that explain the differences in implementation (Schofield, 2001). The organization's perceived needs and benefits are important to whether the implementation of a policy or a program is successful. As suggested by Durlak and DuPre (2008), organizations are more likely to implement a program effectively when they recognize a specific need for a program or policy and believe the policy will yield the desired benefits. It is likely that the successful implementation of a new program or education reform will only be possible if it meets the need for the targeted organization (Kallestad & Olweus, 2003; Tichnor-Wagner et al., 2018). Another organizational factor affecting implementation is an organization's capacity-building conditions, including resources allocated (Durlak & DuPre, 2008; Fixsen et al., 2009; Hatch, 2001) and an organization's starting capacity (Honig, 2006). Examples of resources for policy implementation are money, materials, and time (Durlak & DuPre, 2008; Fixsen et al., 2009; Hatch, 2001; Tichnor-Wagner et al., 2018). Successful implementation of a policy also depends on the starting capacity of the organization or current performance relative to the goal (Honig, 2006). Other contextual factors that have been found to affect implementation are self-efficacy (Durlak & DuPre, 2008), possession of requisite skills (Fixsen et al., 2005; Stith et al., 2006), the degree to which an organization

incorporates a new program into its current practices (Stith et al., 2006), a history of implementing other new programs (Coburn, 2001; Spillane, Reiser, & Reimer, 2002), leadership (Coburn, 2001; Durlak & DuPre, 2008; Hatch, 2001), and effective collaboration and communication (Durlak & DuPre, 2008; Hatch, 2001; Spillane et. al, 2002).

The people in the process of implementation, or actors, were found to constitute another factor affecting policy implementation. Prior research has found procedures that involve clear task-related roles and responsibilities to be important elements of successful implementation (Durlak & DuPre, 2008). Effective leadership and program champions, especially those who are trusted by the other staff in the organization and who can support the new program or policy, have also been recognized as effective boosters for implementation (Durlak & DuPre, 2008; Fixsen et al., 2005; Stith et al., 2006). Sometimes, the same person can serve in both the leadership and the program champion roles (Durlak & DuPre, 2008). According to Honig (2006), previous research usually differentiated among implementers using their formal professional affiliations (e.g., teacher or principal) and assumed that individuals in these roles possessed different backgrounds that configured their involvement in implementation. However, recent research tends toward deeper examination of the implementers' functional roles (Honig, 2006).

Conceptual Framework

Honig (2006) indicates that prior implementation research generally identified policies, people, and places as the three dimensions of education policy implementation. She suggests an interactive framework of contemporary education policy implementation (Honig, 2006). The interactive effects of the three dimensions make implementation a highly situated process, and it is not possible to properly understand the benefits or limitations of one dimension separately from the others (Honig, 2006). This framework also demonstrates the process of implementation

as a multilevel interdependent system that considers the communications and collaborations among the government, organization, community, and other stakeholders in implementation (Honig, 2006).

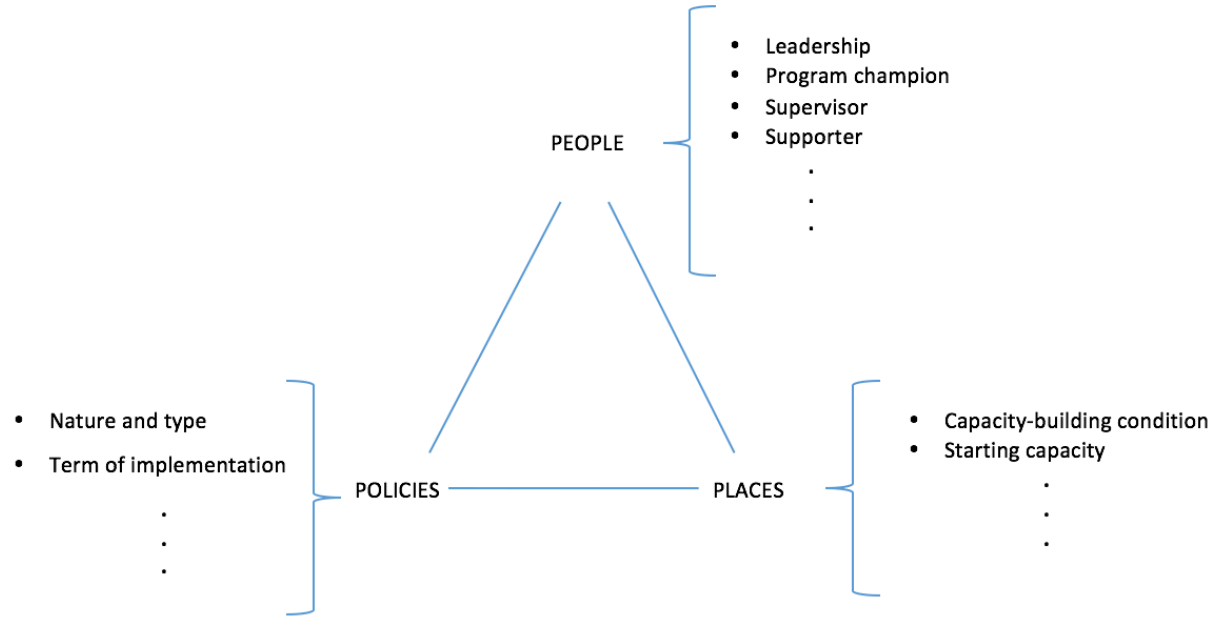

Figure 2. Factors Affecting Implementation in Three Dimensions

Based on the literature review on implementation factors and adapted from the framework that Honig proposed, the conceptual framework of this thesis is presented as a three-dimensional system with policies, places, and people (Figure 2). Each of the policy implementation factors that can be found or quantified in this study's data sources falls into one of the three dimensions. In this research, the conceptual framework focuses only on the level of the organization. This thesis does not have information about the different roles of people in the implementation process. However, the researcher could identify whether a task was assigned to a specific person or group of people who led, supervised, supported, advocated for, or implemented the task, or if it was assigned to a group of people who were not specifically identified. Although the author was unable to examine all factors found in the literature review

that influence policy implementation, this thesis strives to empirically assess the associations between many of these factors and the completion of tasks.

Research Design and Methodology

This study used two sets of data: the text data of planning reports and the administrative data of schools. Measures were extracted from both data sets. This study used two text analysis methods to extract information from the planning reports: the topic model and NER. Then the measures were tested in a multiple logistic regression to find out their association with the outcome variable, whether the task was completed or not.

Data Sources and Data Collection

Data used in this study were publicly available school-level state administrative data and the CSIPIRs.

Text Data

This thesis studies the completion of school improvement tasks reported by the schools seeking improvement. As mentioned previously, these tasks were reported to the state by the identified reforming schools. Most schools submitted the reports three times per year to the state, with descriptions of the tasks planned and to whom the tasks were assigned, until the schools were no longer identified. The schools also reported whether and when the tasks were completed. Thus, the reports possess a longitudinal nature. However, because this thesis focuses on studying whether tasks were completed after the reform period ended, the text data cleaned for this research treat multiple entries of a task as one, and the completion status of each task in the final year was extracted. Thus, the cleaned dataset of tasks used in the analysis does not have a time-series feature.

There are, in total, 43,819 meaningful tasks submitted by the 385 reforming schools over the five years from 2011-12 to 2015-16. There are three types of reform schools that the state identified: SIG, Priority, and Focus. All SIG schools have at least three consecutive years of intervention, while Priority and Focus schools receive funding until they improved their performance and were no longer identified as the lowest performing schools. Schools can be identified again if their performances were later found to be among the lowest. The total number of tasks a school plans is not limited. The total number of tasks ever reported by a school ranges from 12 to 377, with a mean of 114. Among all the tasks, 55.19% were completed. For each school, the percent of tasks completed ranged from 0% to 100%, with an average of 52.25%. The rate of incomplete tasks is high, which shows the necessity of exploring the factors associated with whether or not the tasks were completed.

This analysis used three parts of data from the reports. The first was whether a task was marked with a completion date at its last appearance in the CSIPIRs as the outcome variable. The second part was the tasks themselves. From the descriptions of the tasks, topic modeling was used to identify the abstract topics that occur underlying all the tasks. In the Measures section, an introduction of the topic modeling method is given. This part also briefly presents how the validation was completed. The last piece of information used from the reports is the required field, "Assigned to," for each task reported. Using this information, I then used a natural language processing method called NER to discern whether a task was assigned to any specific person or to a general school agency. Despite the rich information reported by schools in CSIPIRs, these reports were self-reported by the schools and were not consequential. Thus, good quality of reporting was not guaranteed.

WA Report Card Data

WA Report Card is a system of state, district, and school-level data gathered and reported by the state and open to the public on the OSPI website. This analysis used seven years (2010-2016) of school report card data to capture school characteristics. Variables include the percentage of students meeting proficiency levels (math and reading combined), demographics of students, schools' education level, student-teacher ratio, and enrollment. Demographics of students include percentages of students eligible for free or reduced-price lunch, bilingual students, students of color, and students receiving special education.

Measures

The characteristics of tasks were extracted from the text data of CSIPIRs using text analysis methods. The measures about school contexts were extracted from the WA school report card data.

Task Completion as Metric of Implementation

Similar research on education policy implementation has used various measures to indicate the levels of implementation such as frequency and reach (Tichnor-Wagner et al., 2018). Because there is no adequate quantitative metric to test the level of implementation given the CSIPIRs formatting, this study used whether a task was marked as completed to serve as an indicator of the status of implementation for reform tasks. This information was transformed into a binary variable, with "1" representing "completed" and "0" representing "not completed."

Abstract Topics Underlying the Tasks

The measures representing the abstract topics underlying the tasks were adopted from the validated results of topic modeling generated in the text-as-data research on school reform processes (Sun, Liu, Zhu & LeClair, 2019). The first step in exploring whether schools are more likely to implement tasks with certain topics is to quantify the topics of the tasks. Instead of the

traditional way of manually classifying the tasks within different topics, topic modeling, an unsupervised machine learning method for text mining, was used to discover the topics mentioned in the collection of CSIPIRs text. This analysis used the same model setting as the Sun et al. paper, but instead of using topic modeling on only completed tasks, this thesis applied the same model on a different set of samples from the reports of all the tasks (Sun et al., 2019).

The topics were discovered using the statistical model and computerized algorithm, Latent Dirichlet Allocation (LDA; Blei, 2012; Blei, Ng, & Jordan, 2003). The contents of the tasks were first cleaned by removing common English stop words (e.g., "and," "on," "the") and punctuation. The words were also stemmed to remove the suffixes and retain their meaning. Then the topic model was implemented using the R package *stm* on all the words and bigrams. Through topic modeling, the latent topics that compose the reform tasks were discovered among the collection of all the tasks. In LDA, a group of words from the collection of all tasks was found that represents the key information in the collection, and this group of words was the topic. In a topic model, the number of topics is chosen by the researcher. (This process and the validation of the topic modeling are further described in the following paragraph.) At this point, the topics are a subset of words with a high probability of co-occurrence. Then in a task, this model assigned each word to a topic. With words assigned to different topics, each task was then assigned with the probability of its words discussing a certain topic.

Thus, the output of the topic modeling consists of two parts. Topic modeling identifies (1) all the topics it identified in which a group of words were presented as a topic and (2) the probability of each task discussing each topic. If a certain Task A is mainly about Topic 1, then the probability of Task A in Topic 1 will be large, and the probability of Task A's presence in other topics will be small. This probability was called *topic loading* of a task on a topic. To

explain this result, Table 1 presents an example of a topic and what probabilities a task has for this topic and for other topics.

[Table 1 Here]

Validation of the topic model's results is essential since researchers make decisions about model sections that have significant influence on the outcome presented (Sun et al., 2019). Since topics given by the LDA model are based only on statistical calculations, the group of words that compose each topic need to be labeled and validated. Two experts with extensive knowledge of school improvement and K-12 education practices worked independently to evaluate the consistency of each topic and assign labels to the set of words. These two experts are the research team members. One expert coder is a graduate research assistant on this project. He was a classroom teacher in the past and worked extensively with school principals on planning and implementing school improvement. In addition to these practical experiences, he is familiar with the literature on school improvement. The other expert coder is a faculty researcher on this project. She works with districts and schools in co-designing school improvement plans using the schools' own data. She has also published in relevant areas. These two expert coders have both practical and theoretical knowledge to assess the consistency of the topics and label them. Both coders first independently read a sample of tasks highly loaded on a given topic, then labeled each topic by referring the common content of these sample tasks. They also rated the coherence of content across topics. The result of this expert labeling was used to validate text analysis results and choose the optimal model from different topic models.

There are three topic model settings that were tested: the 15, 20, and 30 topics models. Inter-rater reliability was tested using Krippendorff's alpha. The alpha for the three topics models ranged from 0.81 to 0.89. Then, based on the coherence ratings, the 20 topics model was

chosen to be used for later analysis because it contained the highest proportion of coherent topics. The experts agreed to give 15 out of 20 topics concrete labels. These 15 topics are the main themes that schools discussed in their reform tasks. Table 2 includes all 20 topics identified by LDA and their corresponding labels if the average coherence ratings were 3 or above. Only the topics with coherence ratings of 3 or above are included in the later analysis since only these topics have meaningful themes that can cast implications on task completion.

[Table 2 Here]

Based on the topic loadings for a specific task, the researcher then generated the categorical variable **primarily_loaded,** indicating which topic the task is primarily about—that is to say, has the largest loading—among the 15 coherent topics. Some tasks have the highest loading on the incoherent topics. In this analysis, their next highest loaded coherent topic was the primarily loaded one. I chose to use the categorical variable rather than the original topic loadings because it is easier to interpret. Interpretation of results from topic loadings could be difficult because the topic loadings of each task add up to 1 and the loading is derived based on the content of each task. Therefore, the loadings also depend, for instance, on the number of words in a task. Thus, the loadings may not be the best measure for differentiating among tasks. The researchers considered the potential concern over tasks that are equally loaded on two or more topics. Although multiple topics could be discussed in one task, the number of tasks equally loaded on multiple topics is small. For example, to be equally loaded on two topics, one task must have two loadings close to 0.5. The data of topic loadings show that only 9.5% of the tasks have at least one topic loading between 0.4 and 0.6. Furthermore, based on some summary statistics on topic loadings, their standard deviations (SD) are much larger than the mean (e.g., Topic 1 with a mean of 0.031 and SD of 0.146), which indicates that the loadings are spread out.

Therefore, it is reasonable to assume that few tasks are equally loaded on two topics and the use of primarily loaded topics does not create a great loss of the information within topic loadings. This categorical variable was then included as an independent variable in the logistic regression to assess whether the main topic of a task is associated with task completion.

Task Assignments

To obtain the measure of whether the task was assigned to any specific person, the method of NER was used to identify a person's name in the "Assigned to" field in the report for each task. The point person or the school agency that took charge of, supervised, or took an important role in the task implementation was reported in this field. Named entity recognition is an NLP tool that can extract information from text data to find and classify named entities into several predefined categories (person, location, organization, date, etc.). In this analysis, this method was used to identify whether each task was assigned to a specific person or people or to a general agency in the school.

In text documents, there are sequences of words known as named entities that specifically refer to terms that can be denoted by proper names, such as persons, places, company names, organizations, and so on. In the school planning reports, each task was assigned either to one or more specific persons or to a general agency in school (e.g., leadership teams, grade-level teams) without clearly stating the point person(s). Given the large total number of 43,819 tasks, the primary researcher determined it would be most efficient to use a computer-assisted method to identify the names in the "Assigned to" field. I used Stanford CoreNLP to implement this technique on the "Assigned to" part of the text data from the CSIPIRs, mainly to identify individuals' names (Finkel, Grenager, & Manning, 2005; Manning et al., 2014).

Any sequence of words recognized as PERSON in the entry for a task shows that the task was assigned to at least one specific person in the school or from coaches assigned by the state. This means there was at least one specific person who took responsibility for implementing, supervising or supporting the completion of the task. Whereas, a task without any identified PERSON entity shows that it was assigned to a general agency. In later analyses, I explored whether a specific person or a general agency was assigned to was associated with the completion of the task. The variable **person_assign** is a binary with "1" representing that at least one person's name was found, while "0" means that no person's name was found. This measure was then added to the logistic regression as an independent variable.

Other Measures from CSIPIRs

Another measure extracted from the CSIPIRs is **target_duration**, meaning the expected length of time needed to complete a task. This measure was calculated by subtracting the task reporting date from the target completion date. This measure captured the perceived difficulty of a task and whether the planned task was designed to have short- or long-term effects.

School Contextual Measures

Demographics of students, student-teacher ratios, and enrollment were averaged across the 2010-2016 school years to give non-repetitive measures associated with each school. These averages over the years can incorporate the changes in school policies and give a general **school context**. School enrollment was used to capture the size of the schools, the levels of resources allocated to them, and the complexity of the organizations. The prior percentage of students meeting proficiency levels in math and reading was calculated for each school using the average across the three years prior to the first year when they were identified by the state. The average prior measure on state testing (**prior_proficiency**) was then added to the logistic regression to

control for the degree of a school's need to reform and the differential initial performance among all the identified schools.

Samples

Two sets of samples were used in the analysis. The first set of analytical samples were all the tasks, featuring 42,560 in total. The other set included all the tasks except the newly introduced ones. The new tasks were identified as having target completion dates later than the end of the 2015-16 school year (i.e., June 3, 2016). When new tasks were excluded, 6,484 tasks were removed from the analytical samples, bringing the total to 36,076. This thesis tests the hypothesis that whether a task was implemented during the whole period of reform is associated with school characteristics and task features. Despite the number of tasks dropped from the samples, it made sense to also exclude the new tasks when analyzing task completion. Schools may not have spent much time and energy on the new tasks and whether they would be completed remained unknown. Thus, including them could bias the relationship of the task features and school characteristics with task completion.

Data Analysis

Simple multiple logistic regression was first used to detect the factors associated with task completion. Given the nested structure of tasks under schools, I also ran multilevel logistic regression to further discriminate the between and within school effects.

Multiple Logistic Regression

Multiple logistic regression was used to explore whether the probability of task completion is associated with school contexts, latent topics of tasks, and task assignment to any specific person. This analysis also aimed to estimate how much the probability of completing a task changed with the measures extracted from the reports and the state administrative data. The

outcome variable in this analysis is whether or not a task was completed (1 = complete, 0 = not complete) up to the latest year when the data was available, which is school year 2015-16. Thus, the outcome y_i for task i follows,

$y_i \sim Ber(P(y = 1|X = x_i))$, where X is a set of explanatory variables.

As displayed in Equation 1, a multiple logistic regression model was utilized to estimate task *i*'s probability of being completed, as a function of which topic a task was primarily loaded on, whether task *i* was assigned to a specific person, the school's prior percentage of students reaching proficiency levels, and school characteristics:

$$logit\,(P(y = 1|X = x)) = \beta_0 + \beta_1(primarily\_loaded) + \beta_2(person\_assign) + \beta_3(target\_duration)$$
$$\beta_4(school\_context) + \beta_5(prior\_proficiency)$$

where β_1 indicates the relative probability for a task to be completed if the task is primarily loaded on a certain topic as compared to Topic 20 as the base category. β_2 represents the change of conditional probability for a task to be completed given that the task was assigned to a specific school or state staff member. β_3 represents the association between the target length of time to complete a task and whether the task was completed. The **school_context** includes the percentage of students who received free and reduced-price lunch, were bilingual, received special education, and were racial/ethnic minority in a school. It also includes a log of school enrollment, student-teacher ratio, and the education level of a school. β_5 shows how the school's average prior academic achievement on state testing is associated with task completion. Lastly, because the analysis was performed on the task level, the standard errors were clustered at the school level to account for the dependence on the tasks nested in the same school.

As indicated in the pairwise correlation table among pairs of school average characteristics (Table 3), some independent variables have high and significant Pearson

correlations. It was expected that the Pearson correlation between the three school average student demographic measures—percentage of students eligible for free and reduced lunch (ρ = -0.711, p = 0.000), percentage of bilingual students (ρ = -0.418, p = 0.000), percentage of racial/ethnic minority students (ρ = -0.567, p = 0.000)—and the prior percentage of students achieving proficiency levels on state tests would be moderately or largely negative. It was also expected that the correlation among the three variables would be significantly positive. Other noteworthy correlations were the moderate positive correlation between student-teacher ratio and school enrollment (ρ = 0.396, p = 0.000) and the negative correlation between the percentage of students receiving special education (ρ = -0.411, p = 0.000). Thus, the problem of multicollinearity was possible. For more reliable estimates of logistic coefficients of school characteristics, the author decided to delete the highly correlated independent variables. The researcher kept the prior percent of students achieving proficiency levels on state tests, controlling for the schools' pre-reform performance and the percentage of students receiving special education, the logged school enrollment, and the student-teacher ratio for the average schools' characteristics over the six years when data were available.

[Table 3 Here]

Multilevel Logit Model

Multilevel logistic regression was also used to further distinguish effects between schools and within schools. One of the most important assumptions of logistic regression is independence among observations. For the dataset used, this assumption was not fully met since multiple tasks belonged to the same school and both task-level and school-level predictors were added into the logistic regression model. Thus, multilevel logistic regression was also

implemented to incorporate the nested structure of the data. The same set of independent variables was used in the multilevel model and each group was an individual school.

Results

There are noticeable differences in tasks and school characteristics between the groups of complete and incomplete tasks, as indicated by Table 4. In total, 9 of 15 topic proportions reveal significant differences between the two groups of tasks. This result indicates that completed tasks are about different reform topics from those of incomplete ones. Furthermore, the differences in schools' characteristics are mostly significant between complete and incomplete tasks. Generally, tasks are more likely to be completed when associated with schools that serve more racial and ethnic minority students, low socioeconomic status students, and bilingual students; have larger enrollment size; and have lower percentages of students achieving proficiency levels on state tests. However, since the magnitudes of the differences are small, the relationship between the differences in topic proportions or school characteristics and task completion status is difficult to see from the descriptives alone.

[Table 4 Here]

Multiple Logistic Regression Result

Multiple logistic regressions were fitted on the two sets of samples previously described in the Samples section. The first set consisted of the samples of all tasks. The other included all the tasks except the newly introduced ones with target task completion dates later than the end of the 2015-16 school year.

Logistic regression results on both sets of analytical samples are shown in Table 5. When all the tasks were included in the samples, Topic 3 (“Use assessment data to inform interventions and instruction”) and Topic 18 (“Administer and use common assessments”) show moderate

effects on task completion. In other words, tasks primarily focused on using assessment data to inform instruction have higher log-odds of being completed (b = 0.186, a 20% increase in odds ratio), compared with tasks primarily loaded on Topic 20 ("schools' administration activities"). Tasks mainly focused on administering and using common assessments show significantly lower log-odds (b = -0.280) of being completed (a 24% decrease in the odds ratio). The target length of time to complete a task also shows a statistically significant negative coefficient on the log-odds. Because the unit of the target length of time is days, this finding can be interpreted as if one task was expected to be completed in 100 more days, the odds of it finally being completed decreases by 18%. Two of the school characteristics show significant effects. The average percentage of students achieving proficiency levels on state testing prior to reform, like the target task completion time, shows a subtle negative effect. Yet, the logged school enrollment has a modest positive effect on the odds of completion (b = 0.148, a 16% increase on the odds ratio).

These results suggest tasks that involved using assessment data to inform interventions and instruction had a higher likelihood of implementation, and tasks that primarily used common assessments were less likely to be completed. In addition, the longer the implementers expect a task to take to be implemented, the less likely it is to be completed. This may be a result of difficult tasks both being more difficult to implement and taking longer to complete. Moreover, large sized schools are more likely to have their planned tasks completed. Although larger enrollment size could increase the difficulty of task implementation in these schools, it also means they receive more funding and support.

After excluding newly introduced tasks, the negative effect of prior percentages of students reaching proficiency levels and the positive effect of school enrollment became unrecognizable. As for the topic that a task was loaded on, the positive effect from Topic 3

became blurred. One possible explanation for the apparent disappearance of these associations is that some newly identified schools were excluded with the new tasks. Thus, the variations among the schools decreased. However, being primarily loaded on Topic 17 ("Align curriculum with common core and other state standards") showed a decrease in the odds of being completed by 23% (b = -0.260). The other effects remained similar to the model of all tasks. This result indicates that, although schools had them planned, tasks related to aligning curriculum with standards were difficult to complete.

[Table 5 Here]

The goodness of fit indices for the two models were acceptable but not satisfactory. As shown in the classification table (Table 6), both the all-tasks and excluding models had correctly classified approximately 60% of the data (60.60% and 64.92%, respectively). The overall fit of both models was comparable. Both models performed well in correctly predicting the completed tasks with sensitivities of 82.85% and 96.81%, respectively. However, both models performed poorly on the incomplete tasks. In the model that excluded new tasks, only 3.79% out of all the incomplete tasks were correctly classified as incomplete. Given the unsatisfactory goodness of fit, the model still needed significant improvement. However, the interpretation of the coefficients was still valid and could give useful implications for school improvement implementation. As the next step of the analysis, instead of a multiple logistic regression, a multilevel logistic regression was fitted with two levels (task level nested in school level) to further detect the effects from predictors at both levels.

[Table 6 Here]

Multilevel Logit Models

First, an unconditional mean model was fitted to both analytical samples to determine whether a multilevel modeling was necessary and to show the decomposition of variance between and within school components. In this analysis, the proportion of variance in the probability for a task to be completed that lies between different schools was estimated. Then each Intraclass Correlation Coefficient (ICC) was computed. The ICCs for the all-tasks and excluding new tasks models were 0.18 and 0.20, respectively, which indicated that 18% or 20% of the variance was explained by the between-school differences. This amount was not trivial and suggested that it was necessary to consider multilevel modeling.

Then, the random intercept model was fitted to both samples. The results are presented in Table 7. Random slope models were not considered in this analysis because there is no prior literature on any task-level characteristics performing differently among schools. Thus, the simpler random intercept model was used.

[Table 7 Here]

The multilevel modeling results in both sets of samples find additional factors associated with task completion. When all tasks were included in the samples, Topic 6 ("Implement teacher evaluation via classroom observation, teacher reflection, and observer feedback") has moderate positive association on task completion. The association indicates that tasks primarily focused on teacher evaluation are more likely to be fully implemented (b = 0.135), compared with those primarily loaded on Topic 20. The odds for these tasks to be completed increases by 14%. In terms of school contextual factors, assigning tasks to a specific person(s) shows positive association with task completion (b = 0.085). If a task is assigned to a specific person instead of a general agency, the odds of completion for this task will increase by about 9%. The positive association between enrollment and task completion was not detected in this model. The

estimated association among Topic 3, Topic 18, prior academic achievement, and target length of time to complete remained at the same level compared with the multiple logistic regression.

After excluding newly introduced tasks, similar to the multiple logistic regression results, the subtle negative effect of prior academic achievement and the positive effect of school enrollment became non-recognizable. The positive effect from Topic 3 and Topic 6 became blurred. Primarily loaded on Topic 17 ("Align curriculum with common core and other state standards") and Topic 18 ("Administer and use common assessments") showed moderate negative association with tasks completion. Assigning tasks to a specific person(s) was another factor found to have significant and moderate positive association with task completion. The magnitude of this factor became larger compared with the all-tasks model ($b = 0.114$). This result suggests that a task assigned to a specific person(s) can expect a 12% increase in its odds of being completed during the reform period.

The all-tasks model and the model excluding new tasks correctly classified 65.44% and 69.26% of the data, respectively, as shown in the classification table (Table 8). Both models performed well in correctly predicting the completed tasks with sensitivities of 78.89% and 90.53%, respectively. Despite slightly lower sensitivities, both multilevel models classified many more incomplete tasks correctly. For the all-tasks model, the specificity increased by about 16%. For the model excluding new tasks, there was an increase of approximately 25% in correctly predicting incomplete tasks. Thus, compared with the multiple logistic regressions, the multilevel models are better at classifying the completion status of tasks.

[Table 8 Here]

Discussion

In this study, characteristics of the school reform tasks were extracted using text analysis methods. These task-specific characteristics, together with school contexts, were tested using logistic regression on their association with task completion. Across the four logistic regression models, the results consistently show associations between task-specific characteristics and task completion. For example, this study indicates that tasks with certain topics are more likely to be completed. Across all the models tested, when compared with schools' administration activities (Topic 20), schools' planned tasks around administering and using common assessments are less seen promised completion, although using common assessments was identified by the topic modeling as one of the main abstract topics. A possible explanation is that, compared with other tasks, using common assessments requires significantly more training on the assessments themselves and how to maintain the efforts related to completing these tasks. Similarly, tasks about Topic 17 ("Align curriculum with common core and other state standards") are less likely to be completed due to the greater need for professional development in the knowledge and understanding of the Common Core State Standards (Smith & Thier, 2017). Interestingly, in both all-tasks models, tasks primarily loaded on the other topic about assessments—Topic 3 ("Use assessment data to inform interventions and instruction")—are more likely to be completed. It is worth noticing that this topic is written by the schools and has an objective and desired outcome. The results lead to a similar conclusion as found in the literature that policies with clear and consistent objectives are more likely to be implemented successfully *(*Berman 1978; Bullock 1980; Ripley and Franklin 1982; Schofield, 2001). Similarly, Topic 6 ("Utilize school leadership teams to drive school improvement"), which shows positive association with completion in the multilevel logit in the all-tasks model, is also written with an objective. However, one limitation

of this finding is that the topics were labeled by the two experts, and the objectives were from interpretations of the topic words.

The targeted length of time for tasks to be completed has a negative association with whether or not the tasks were completed. The researcher expected to see that tasks perceived to be more difficult by the schools were less likely to be completed at the end. The other factor extracted from the report, the assignment of tasks to a specific person versus a general agency, has a positive association with task completion as identified in the multilevel models. This result suggests that assigning a planned task to a specific person(s) helps the implementation of the task and increases the probability of the task's completion. This finding is consistent with the prior literature concerning the way support from specific staffing designations such as leader, program champion, or supervisor impacts implementation on a large scale (Durlak & DuPre, 2008). The opposite of assigning to a specific person in the school improvement reports is usually a group in a school (e.g., grade level teams, math teachers). Assigning to the group does not necessarily guarantee a good level of collaboration among group members, although prior research has identified collaboration as an important factor in implementation. The finding in this study suggests that assigning a go-to person(s) to lead, support or supervise a task facilitates task completion.

In summary, the results find support for the association between the characteristics of the tasks themselves and their completion status. The results also cast new light on the usefulness of text analysis methods as ways to use school reports to reveal policy implications. Exploration of task characteristics in this thesis relied on the text analysis methods including topic modeling and NER. Topic models allowed the researchers to uncover the abstract topics underlying the tasks. The NER provides an efficient way to identify the go-to person(s) assigned to each task. With

these methods, the computer assists the process of reading and labeling text. These methods are especially useful when it is difficult to analyze a large amount of text manually. However, these methods have their limitations. For example, despite recognizing the person name(s), an application of NER without additional features does not allow the primary researcher to identify the assigned-to persons written as their titles, such as principal, leader, and chair. This limitation does not distort the results of this thesis since the proportions of these titles are small (1.14%, 0.06%, and 0.04% for "principal," "leader" and "chair," respectively). However, this could be a problem when the proportions are high. Moreover, results from these text analysis methods depend on the quality of text data. Relatively low-quality reports may distort the results and make the results hard to interpret. The results, however, indicate an unexpected null association between most school contexts and the completion of tasks. Enrollment is the only school characteristic that has a moderate positive association with the completion of tasks, but the relationship was found only in the all-tasks models. In general, the association between school contexts and task completion was not consistently found from the results. Contrary to prior studies, this analysis does not support the idea that school contextual factors are associated with task completion. However, most of the school contextual variables used in this analysis were averaged over seven school years (2010-2016). Accordingly, variations of these contextual factors over the time period of 2010-2016 are not considered in the models. Also, the average over the years can be affected by extreme values if the schools encounter changes in policy, leadership, or school climate. Using averages across years also makes the pre-reform and post-reform effects inseparable. School contextual factors might be confounded with the tasks themselves. This may also explain the lack of associations between the completion of tasks and school contextual factors.

Based on earlier research, I also expected to find an association between the schools' academic performance prior to reform and the completion of tasks. Prior literature has indicated that the starting capacity or current performance relative to the goal is an important factor affecting implementation. However, the results of this analysis do not find such an association between schools' prior academic performance and task completion. The most reasonable explanation of this result is that the prior academic performance of schools may not have been greatly varied because all the schools that implemented the WA school improvement policy are low-performing schools. Even though the reforming schools' starting academic performances were different, the variation may not have been large enough to impact task implementation. Moreover, it is reasonable to believe that there are factors other than school characteristics that have a direct impact on task implementation.

This thesis has several limitations. First, the only available outcome measure for implementation is whether or not the task was completed. If future research can find other data to quantify implementation, the association with prior academic performance could possibly be detected. Future research can also explore whether starting academic performance affects student academic outcomes after a policy is implemented when the two measures are more directly related. A second limitation is that the logistic regression ignores the longitudinal nature of the data. These longitudinal data sets should be explored by future studies, which can identify further insights into the process of implementation, such as when each task was completed and how tasks were implemented over time. The third limitation of the data for this thesis is the limited description of the assigned-to persons. Without knowing their roles in the implementation process, this study was only able to contrast specific person to general agency rather than to contrast the different roles of each person. If this type of data is reflected in the report,

researchers can use it to further understand the process of implementation. This research also has limitations in testing all the implementation factors found in the literature review due to unavailable data on these measures. This is an exploratory study on task completion and does not have the power to draw causal claims. The last limitation is from the quality of the CSIPIRs. Although I was able to generate measures from the text data using text analysis methods, the quality of reporting limited the results. Since the CSIPIRs are not consequential, the reported contents may not be accurate and consistent. Therefore, the results should be interpreted carefully.

Conclusion and Implications

This thesis suggests that the completion of reform tasks is associated with the topics on which tasks were primarily loaded and to whom the tasks were assigned. However, the null associations between schools' prior academic performance and other contextual factors from the logistic regressions are unexpected. These findings may have implications for state and school practitioners in the implementation of school reform policy. Specifically, since some task-specific characteristics are found to have a moderate association with task completion, school practitioners should pay extra attention to setting up reform tasks with clear goals and assigning each reform task to a specific person rather than a general agency or no one. Schools should also address the importance of keeping reliable and detailed reporting so that the state can use these reports to review schools' reform practices and support successful implementation. Researchers can study the process of implementation using the reports and advance the knowledge base of implementation literature.

Moreover, this thesis can serve to alert state practitioners to the fact that consistent training and professional development may make the implementation of certain tasks more

effectively. State practitioners can also glean from this analysis that it may be more difficult to fully implement some planned tasks that fall within certain topics. The nature of the tasks may be a factor affecting implementation. Practitioners could start with this fact as they continue to explore why these tasks are less likely to be completed. As the null association between most school contextual factors and task completion suggests, the completion of reform tasks as an outcome of implementation of school reform plans depends more on what types of tasks the schools plan and how the schools ensure the tasks' implementation, rather than schools' initial academic performance or other characteristics. Thus, extra effort and support should be given to the implementation process, no matter what type of schools are being reformed.

For policy implementation researchers, this thesis provides an example of applying text analysis methods on a large scale to extract useful information from school reports. Researchers may further explore the possibility of using other text about implementation from actors of different levels and applying other text analysis methods, such as part-of-speech tagging and sentiment analysis, to study language use in the text data.

Given the limitations in the available data about the degree of implementation and the process of implementation, future research could find other ways to quantify implementation and explore the factors affecting implementation from a causal perspective. Future research may also find ways to quantify other implementation factors that this thesis was not able to measure. For example, later research can gather information about the different roles of school staff in the implementation process and data about resources received, such as budget, training, and technical support, so as to probe into the factors that affect the implementation process.

References

Berman, P. (1978). The study of macro and micro implementation. *Public Policy*, **27**, 157–184.

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.

Bullock, C.S. (1980). Implementation of equal education opportunity programs: a comparative analysis. In Mazmanian, D. and Sabatier, P.A. (eds), *Effective Policy Implementation.* Lexington, MA: Lexington Books.

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

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American journal of community psychology*, *41*(3-4), 327-350.

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice, 19*(5), 531–540.

Fixsen, D., Naoom, S., Blase, K., Friedman, R., Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature*. Tamps, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network.

Hatch, T. (2001). Incoherence in the system: Three perspectives on the implementation of multiple initiatives in one district. *American Journal of Education, 109*(4), 407–437.

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

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.

Honig, M. I. (Ed.). (2006). *New directions in education policy implementation: Confronting complexity*. Albany, NY: State University of New York Press.

Finkel, J. R., Grenager, T., & Manning, C. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 363-370). Association for Computational Linguistics. Retrieved from http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

Kallestad, J. H., & Olweus, D. (2003). Predicting teachers' and schools' implementation of the Olweus bullying prevention program: A multilevel study. *Prevention & Treatment*, *6*(1), 21a.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60). Retrieved from https://www.aclweb.org/anthology/P14-5010

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.

Ripley, R.B. and Franklin, G.A. (1982). *Bureaucracy and Policy Implementation.* Chicago, IL: Dorsey Press.

Schofield, J. (2001). Time for a revival? Public policy implementation: a review of the literature and an agenda for future research. *International Journal of Management Reviews*, *3*(3), 245-263.

Smith, J., & Thier, M. (2017). Challenges to Common Core State Standards Implementation: Views From Six States. *NASSP Bulletin*, *101*(3), 169-187.

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

Stith, S., Pruitt, I., Dees, J., Fronce, M., Green, N., Som, A. et al. (2006). Implementing community-based prevention programming: A review of the literature. *Journal of Primary Prevention*, 27, 599–617.

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., Liu, J., Zhu, J. & LeClair, Z. (2019). *Using a Text-as-Data Approach to Understand Reform Processes: A Deep Exploration of School Improvement Strategies.* (EdWorkingPaper: 19-68). Retrieved from Annenberg Institute at Brown University: http://www.edworkingpapers.com/ai19-68

Tichnor-Wagner, A., Allen, D., Socol, A. R., Cohen-Vogel, L., Rutledge, S., & Xing, Q. W.. (2018). Studying Implementation Within a Continuous-Improvement Process: What Happens When We Design With Adaptations in Mind? *Teachers College Record*, *120*(5).

Figure 1. Examples of Reform Tasks Reported in CSIPIR

3. The STAR Enterprise assessment suite has been purchased and is being used by teachers to collect student growth data. The assessments will take place approximately every three weeks for the remainder of the year to monitor progress in math and language arts and gather data on the common core state standards.

	Assigned to:	Tammy Swanson
	Added date:	01/23/2014
	Target Completion Date:	01/21/2014
	Task Completed:	2/18/2014 12:00:00 AM

3. Student learning portfolio development will include individual student goal-setting based on fall assessments. Students will continue to develop their portfolios, in partnership with teachers, throughout the year. Essentially these portfolios work as individualized education plans for each student.

	Assigned to:	All teachers
	Added date:	10/15/2013
	Target Completion Date:	06/13/2014

Note. The first example is a completed task with a marked task completion date that is assigned to a specific person. The name of the person was made up. The second example is a task that has not been completed. There is no task completion date reported. The second task is assigned to a general agency rather than a specific person.

Table 1. Example of Raw Topic Loadings

School Id	Task Id	Topic 1	Topic 2	Topic 3	.....	Topic 18	Topic 19	Topic 20	Sum of Topic Loadings
291002946	532608	0.00213	0.90866	0.00356		0.00449	0.00163	0.00215	1
291002946	100446	0.00222	0.00713	0.00427	.....	0.00376	0.00200	0.03062	1
212143112	146648	0.00226	0.00586	0.01406		0.89787	0.00209	0.00286	1
212143112	447823	0.00211	0.00603	0.01207		0.90934	0.00194	0.00264	1

Table 2. Topic Labels and Coherence Ratings

Topics	Mean Coherence Rating
1. Develop evidence-based student support strategies and interventions	4
2. Schedule and plan a variety of things	1
3. Use assessment data to inform interventions and instruction	4
4. Wrap-around services	1
5. Utilize school leadership teams to drive school improvement	3
6. Implement teacher evaluation via classroom observation, teacher reflection, and observer feedback	4
7. Grade-level team collaborative activities (e.g., reviewing data, co-planning, establishing standards, and principal/AP participation in team activities)	4
8. Professional development on effective teaching frameworks (Sheltered instruction; 5D, Danielson; individual determined; cultural competent)	4
9. School leadership and student supports for socio-emotional, safety, behavioral, and academic progress	2
10. Teachers implement effective instruction and administrative supports for the implementation	3
11. Implement Positive Behavioral Interventions and Supports (PBIS)	4
12. Develop, support, and monitor teacher professional learning communities for instructional improvement	4
13. Engage and communicate with parents and families	4
14. Train staff on guided-language-acquisition-design and other instructional topics	3.5
15. Support students' transition to the next educational level and others	2
16. Provide ELL or SPED students with targeted support	3
17. Align curriculum with common core and other state standards	4
18. Administer and use common assessments	4
19. Goal setting and planning	1
20. School administration activities and support for struggling students	3

Table 3. Pairwise Correlations among School Contextual Factors

	1.	2.	3.	4.	5.	6.	7.
1. Prior % proficient	-						
2. Mean % eligible for free or reduced-price lunch	-0.711***	-					
2. Mean % bilingual	-0.418***	0.654***	-				
4. Mean % special education	-0.149***	-0.023***	-0.090***	-			
5. Mean % racial/ethnic minority	-0.566***	0.737***	0.678***	-0.029***	-		
6. Logged enrollment	0.288***	0.047***	0.321***	-0.121***	0.232***	-	
7. Mean student teacher ratio	0.175***	-0.190***	0.090***	-0.411***	-0.039***	0.396***	-

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

Table 4. Comparison of task and school characteristics between complete and incomplete tasks

		Complete	Incomplete
Average Topic Proportions	Topic 1	0.030 (0.146)	0.031 (0.147)
	Topic 3*	0.060 (0.191)	0.052 (0.171)
	Topic 5*	0.081 (0.204)	0.076 (0.195)
	Topic 6*	0.064 (0.203)	0.058 (0.191)
	Topic 7	0.054 (0.173)	0.0556 (0.176)
	Topic 8	0.047 (0.177)	0.046 (0.177)
	Topic 10*	0.032 (0.149)	0.028 (0.138)
	Topic 11*	0.058 (0.200)	0.064 (0.211)
	Topic 12*	0.053 (0.186)	0.058 (0.194)
	Topic 13*	0.0934 (0.255)	0.087 (0.245)
	Topic 14	0.063 (0.182)	0.061 (0.180)
	Topic 16	0.044 (0.172)	0.043 (0.168)
	Topic 17*	0.055 (0.179)	0.065 (0.196)
	Topic 18*	0.045 (0.159)	0.062 (0.190)
	Topic 20	0.039 (0.159)	0.040 (0.165)
Average School Characteristics	% racial/ethnic minority*	62.557 (25.985)	59.753 (26.860)
	% eligible for free or reduced-price lunch*	72.767 (17.247)	71.232 (18.511)
	% bilingual*	21.116 (18.973)	19.163 (18.238)
	% special education	13.983 (5.785)	13.891 (7.115)
	Prior % proficient*	46.860 (11.782)	47.516 (12.277)
	Student teacher ratio*	16.003 (3.257)	16.083 (3.257)
	Enrollment*	501.664 (294.107)	485.395 (283.808)

Note. * indicates statistically significant at the 0.05 level for the two sample t-test. Only coherent topics are included. School characteristics are averaged over six years (2010-11 through 2015-16), except that prior % proficient is the three-year average prior to reform.

Table 5. Multiple Logistic Regressions Results for Task Completion

	All Tasks		Excluding New Tasks	
	M	*(SE)*	*M*	*(SE)*
Prior percent proficient	-0.007*	(0.003)	-0.002	(0.004)
% special education	0.007	(0.009)	-0.003	(0.011)
Logged enrollment	0.148*	(0.067)	0.111	(0.069)
Student teacher ratio	-0.009	(0.018)	-0.011	(0.017)
Secondary School	-0.072	(0.067)	-0.092	(0.080)
Primarily loaded on Topic 1	0.010	(0.084)	-0.020	(0.092)
Primarily loaded on Topic 3	0.186**	(0.070)	0.089	(0.081)
Primarily loaded on Topic 5	0.062	(0.060)	-0.001	(0.068)
Primarily loaded on Topic 6	0.124	(0.066)	0.032	(0.075)
Primarily loaded on Topic 7	0.005	(0.066)	-0.072	(0.076)
Primarily loaded on Topic 8	0.037	(0.061)	0.035	(0.073)
Primarily loaded on Topic 10	0.076	(0.083)	0.009	(0.093)
Primarily loaded on Topic 11	-0.044	(0.067)	0.062	(0.077)
Primarily loaded on Topic 12	-0.068	(0.074)	-0.049	(0.083)
Primarily loaded on Topic 13	0.106	(0.058)	0.063	(0.069)
Primarily loaded on Topic 14	0.055	(0.064)	0.015	(0.073)
Primarily loaded on Topic 16	0.051	(0.075)	-0.057	(0.084)
Primarily loaded on Topic 17	-0.106	(0.062)	-0.260***	(0.072)
Primarily loaded on Topic 18	-0.280***	(0.068)	-0.276***	(0.077)
Task assigned to specific person(s)	-0.014	(0.067)	0.022	(0.069)
Target length of time to task completion	-0.002***	(0.000)	-0.001***	(0.000)
Constant	0.162	(0.359)	0.541	(0.398)
N (observations)	42560		36076	

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

Table 6. Classification Table for Results of Multiple Logistic Regressions

	All Tasks	Excluding New Tasks
Sensitivity	82.85%	96.81%
Specificity	32.64%	3.79%
Correctly classified	60.60%	64.91%

Table 7. Multilevel Logistic Regression Model Results for Task Completion

	All Tasks		Excluding New Tasks	
Fixed Effects	*M*	*(SE)*	*M*	*(SE)*
Prior percent proficient	-0.008*	(0.004)	0.001	(0.004)
% special education	0.016	(0.011)	0.007	(0.013)
Logged enrollment	0.131	(0.070)	0.102	(0.084)
Student teacher ratio	-0.004	(0.014)	-0.009	(0.016)
Secondary School	-0.127	(0.085)	-0.147	(0.103)
Primarily loaded on Topic 1	-0.030	(0.074)	-0.044	(0.084)
Primarily loaded on Topic 3	0.156*	(0.064)	0.097	(0.073)
Primarily loaded on Topic 5	0.049	(0.058)	0.020	(0.067)
Primarily loaded on Topic 6	0.135*	(0.063)	0.074	(0.071)
Primarily loaded on Topic 7	0.017	(0.065)	-0.021	(0.074)
Primarily loaded on Topic 8	0.064	(0.066)	0.102	(0.076)
Primarily loaded on Topic 10	0.072	(0.075)	0.027	(0.085)
Primarily loaded on Topic 11	-0.004	(0.062)	0.086	(0.072)
Primarily loaded on Topic 12	-0.016	(0.064)	0.007	(0.074)
Primarily loaded on Topic 13	0.099	(0.056)	0.068	(0.064)
Primarily loaded on Topic 14	0.059	(0.062)	0.075	(0.071)
Primarily loaded on Topic 16	-0.028	(0.068)	-0.089	(0.077)
Primarily loaded on Topic 17	-0.053	(0.063)	-0.149*	(0.071)
Primarily loaded on Topic 18	-0.298***	(0.065)	-0.284***	(0.074)
Target length of time to task completion	-0.002***	(0.000)	-0.001***	(0.000)
Task assigned to specific person(s)	0.085*	(0.034)	0.114**	(0.038)
Constant	0.030	(0.448)	0.306	(0.542)
Random Effects	*Var*		*Var*	
Schools	0.551		0.807	
N (Tasks)	42560		36076	
N (Schools)	363		363	

Note. Maximum Likelihood estimates shown. * p <.05, ** p <.01, *** p <.001.

Table 8. Classification Table for Results of Multilevel Logistic Regressions

	All Tasks	Excluding New Tasks
Sensitivity	78.89%	90.53%
Specificity	48.86%	28.69%
Correctly classified	65.44%	69.26%