Title

Evidence for the Feedback Role of Performance Measurement Systems

Permalink
https://escholarship.org/uc/item/490136rc

Journal
Management Science, 65(9)

ISSN
0025-1909

Authors
Anderson, Shannon W
Kimball, Amanda

Publication Date
2019-09-01

DOI
10.1287/mnsc.2018.3129

Evidence for the Feedback Role of Performance Measurement Systems

SHANNON W. ANDERSON (*)
University of California, Davis; One Shields Ave.; 3414 Gallagher Hall
Davis, CA 95616
Email: swanderson@ucdavis.edu

AMANDA KIMBALL
University of California, Davis; One Shields Ave.
Davis, CA 95616
Email: alkimball@ucdavis.edu

FINAL VERSION

* Corresponding author. We are grateful to senior managers and teachers at ASPIRE Public Schools and at Schoolzilla for providing data access and giving generously of their time to help us understand their use of performance measurement system data. The paper has benefited from workshops at the University of Melbourne, the University of New South Wales, Sydney University, Oregon State University, the Naval Post-graduate School, the University of Pittsburgh, Bond University, Bocconi University, and University of Southern Denmark. We are particularly grateful to Harry Evans, Pedro Ferreira, Steven Hansen, Nandu Nagarajan, Sujay Nair, Mina Pizzini, Adam Presslee, Zhiru Sun, Emmile Brack, and Leo Bialis-White for comments on early drafts of the manuscript.

Evidence for the Feedback Role of Performance Measurement Systems

Abstract

Performance measurement systems (PMSs) are used to diagnose and remediate problems, termed the 'decision-facilitating' or feedback role of management control. We examine whether use of PMSs by individual decision makers is associated with better performance. Experimental studies have isolated individual-level effects of feedback on decision quality; however, it is difficult to extend these findings to natural settings. Archival and survey studies offer evidence on the association between the *presence* of PMSs and performance, but have had limited success in measuring decision makers' actual *use* of PMSs and in addressing endogeneity of the decision to use PMSs. We use unobtrusively collected data on actual PMS use in 30 K-12 charter schools over three years to test whether teachers who make greater use of two PMSs are associated with greater growth in student learning. We find that teachers' use of PMSs is associated with increased student learning, consistent with the premise that PMSs facilitate teacher interventions and improve student outcomes. The results are both statistically and materially significant and are better explained by PMS use than by selection effects of better teachers using PMSs. Consistent with the organization's focus on "at risk" students, the strongest effects of teachers' use of one PMS are concentrated among the lowest-performing students. In sum, we find broad support for the thesis that the feedback role of PMSs is associated with meaningful performance improvement.

Key Words: performance management, management information system, education sector, assessment

Data Availability: The data are the property of Aspire Public Schools and may not be redistributed by the authors.

Evidence for the Feedback Role of Performance Measurement Systems

1. Introduction

Performance measurement systems (PMSs) are widely used to provide decision makers with routine access to multifaceted performance assessments, framed in relation to strategic objectives and performance goals. The performance measurement dashboards and scorecards popular for decades in for-profit companies have been adapted to the education sector in response to demands for accountability from politicians and from philanthropic "education entrepreneurs" with roots in the corporate world (Visscher and Coe 2003, Ikemoto and Marsh 2007). However, Staman, Timmermans and Visscher (2017, 59) note, "despite high expectations of DBDM [data-based decision making] in terms of improved student performance, scientific evidence on the effectiveness of DBDM interventions is still rather limited."

In business, PMSs typically serve two purposes: to evaluate and reward results, termed the "decision-influencing" or incentivizing role; and to diagnose problems and devise interventions aimed at improving performance, termed the "decision-facilitating" or feedback role (Demski and Feltham 1976). In the education sector, incentive pay for teachers has been shown to have limited impact on student learning outcomes (Fryer 2013, Firestone and Pennell 1993). In the K-12 charter school organization that we study, incentive pay is not used for either teachers or school principals. Thus, this study isolates the decision-facilitating role of PMSs in the organization to examine empirically whether greater use of PMSs is associated with better learning growth. We use three identification strategies to distinguish the posited "feedback effect" of PMS use from a "selection effect" of better teachers using PMSs.

The PMSs that we study supply teachers with daily feedback on student learning. They support trend analysis and other meaningful comparisons, and are consistent with models of effective feedback and decision support (e.g., Keuning, Van Geel and Visscher 2017, Stecker, Fuchs and Fuchs 2005). Both PMSs use data visualization (Cardinaels and Van Veen-Dirks 2010) to direct teachers' attention to students' shortcomings against learning objectives, to cue the need for remediation, and to support lesson planning to execute the remediation strategy. Based on theory of effective feedback systems, we hypothesize that teachers who make greater use of the PMSs are associated with students who have greater learning achievement and that this feedback effect is present in addition to any selection effect.

Although the PMSs were adopted by the organization with the expectation of positive effects and their use by teachers is voluntary, a positive and economically meaningful feedback effect of teachers' use of the PMSs on student learning is far from assured. PMS use may fail to yield appreciable improvement in learning outcomes if: 1) PMS use comes at the expense of more effective teaching activities, 2) PMS use is intended to support other valued outcomes (e.g., prosocial behavior) that may be weakly related to learning, or 3) the PMSs confirm what the teacher already knows, providing no incremental information (Keuning et al. 2017). Moreover, as Hattie (2009) documents in a comprehensive review of hundreds of meta-analyses of thousands of published intervention studies, PMSs, like many education interventions, may exhibit a statistically significant positive association with learning outcomes but at an effect size that is too small to warrant policymakers' attention.

Management accounting research generally supports a positive association between PMSs and organizational performance. However, common limitations are: 1) a failure to distinguish the feedback effect of PMS use from the roles of incentives or selection effects, 2) a failure to link demonstrated effects on decision makers to organizational outcomes, or 3) statistical associations that do not translate into meaningful effects on organizational performance. For example, experimental studies that isolate the feedback role from selection effects and incentive effects find that PMSs have a causal relation with individual-level decision making (e.g., Lipe and Salterio 2000; Libby, Salterio, and Webb 2004; Vera-Munoz, Shackell, and Buehner 2007; Taylor 2010; Cheng and Humphreys 2012). However, these studies typically do not extrapolate individual-level effects to organizational outcomes. Moreover, participants' use of the PMS as prompted by the experimental task, may not mirror choices about PMS use in a natural setting, and effect-size cannot be evaluated because it is determined by the strength of the experimental manipulation (Sprinkle 2003). Organizational outcomes are more readily measured in archival field-based research, but many studies do not directly measure decision makers' use of PMSs, and cannot separate the feedback effect of PMSs from either selection or incentive effects (e.g., Ittner, Larcker, and Randall 2003; Davis and Albright 2004, Casas-Arce, Martinez-Jerez, and Narayanan 2017).

To examine the association between the feedback effect of PMSs on decision makers and its association with organizational performance, this study capitalizes on a novel setting and modern information technologies to bridge gaps in the literature that have at their roots the comparative strengths and weaknesses of different empirical methods. Specifically, in a natural setting that is relatively uninfluenced by incentive pay we: 1) measure actual, unprompted PMS use by a decision maker in a natural setting; 2) measure actual performance outcomes that the PMS is intended to influence; 3) evaluate both statistical significance and materiality (effect size) of the association between PMS use and performance; 4) exploit time series properties of the data to disentangle the effects of PMS use and user selection effects; and 5) test whether the association between PMS use and performance is consistent with the decision maker using PMSs in a strategy-directed manner to achieve specific organizational goals.

The research site for this study is ASPIRE Public Schools (ASPIRE), a large K-12 charter school system that uses two PMSs to support teachers' efforts to improve student performance on standardized statewide tests.[1] The PMSs supply teachers with daily and weekly feedback throughout the school year on students' mastery of key learning objectives with comparative benchmarks at the student, class and school levels. As PMSs have migrated to internet-based platforms that can record and archive user keystrokes, these technologies create new opportunities for measuring actual PMS use in the field. Neither the PMS usage data nor student test outcomes are used explicitly in teacher evaluation; indeed, we collect the recorded measures of teachers' actual daily use of the PMSs specifically for this study, as teacher PMS use has never been monitored by the organization. In many PMSs, feedback is intended to facilitate strategy execution. Field interviews confirm that teachers use the data for lesson planning and to target interventions for students who do not demonstrate mastery of learning objectives.

[1] ASPIRE is one of 20 school systems described in McKinsey & Co. (2010) and is widely acknowledged to be a well-managed organization that is at the forefront of successful education reform.

We test for a positive association between teachers' actual daily use of the PMSs during the year and students' end-of-year academic performance (i.e., learning growth in relation to annual standardized test outcomes). We use three years of data and three identification strategies to disentangle the hypothesized feedback effects of the PMSs from the selection effects of better teachers choosing to use the PMSs. Although the chronology of PMS use preceding performance outcomes is insufficient to prove causality, it is a necessary condition for testing theory that PMSs provide feedback that facilitates actions that improve performance. Like many charter schools, ASPIRE's focus is on accelerating growth among "at risk" students. Thus, a more refined test of the research question is whether the association between PMS use and performance is stronger for at risk students. In sum, the organization provides a good quasi-experimental setting in which to estimate feedback effects of PMSs.

Overall, we find that teachers' use of PMSs creates a feedback effect that is associated with increased student growth in both English language arts (ELA) and math. For one of the two PMSs, there is further evidence that teachers use it in a strategy-directed manner to focus on at risk students. In particular, the PMS that was custom-designed for the organization exhibits a positive association between teachers' PMS use and math growth that is much stronger in the lower deciles than in the upper deciles of realized student growth. The second PMS, which is purchased from an external vendor, exhibits a uniform positive effect for all levels of realized student growth.

The interpretation of these results as evidence of a feedback effect rather than a selection effect is premised on three identification strategies. First, the association holds after controlling for teacher effectiveness, as evaluated by the teacher's principal. Second, we use a difference-in-differences model in which the teacher serves as her own control to demonstrate a continued effect of PMS use after removing teacher effects. Finally, we use simulation analysis to identify the distribution of the correlation between contemporaneous performance and PMS use when the correlation is caused by teacher selection alone. The actual observed correlation between performance and PMS use falls in the far-right tail of the simulated distribution, indicating that the estimated effects are more likely to be explained by feedback effects of PMSs than by selection effects. External validity is provided by the finding that teachers' use of

PMSs has a stronger association with math than with ELA growth. This accords with prior studies that find teachers have a greater influence on math than on ELA test scores (Rich 2013; Tuttle, Gill, Gleason, Knechtel, Nichols-Barrer, and Resch 2013).

The estimated feedback effects of PMS use are both statistically and materially significant. To estimate effect size, we define three "treatment effects" using teachers' above-median use of: both PMSs, the first PMS only, and the second PMS only. We compare student outcomes for these teacher groups to those of teachers with below-median use of both PMSs. For math, all three treatments are significant, and for ELA the treatments of above-median use of both PMSs and the second PMS only are significant. The effect sizes that we measure for PMS use compare favorably with effect sizes of other education innovations that are judged to be material for informing policy (Hattie 2009). We demonstrate the external validity of the effect size by showing that the effect sizes of exposing students to teachers with above- versus below-median effectiveness are comparable to those of prior studies.

In summary, the study contributes important evidence to the management accounting literature on the decision-facilitating role of PMSs. We use a novel setting to link real decision-makers' actual *use* of PMSs to organizational performance, finding that the predicted positive association is statistically and materially significant. Importantly, we demonstrate that the performance improvements associated with PMS use are distinct from selection effects associated with decision to use the PMS.

The paper is organized into six sections. Section 2 reviews the research literature on feedback and its theorized effects on performance, with a focus on feedback supplied through PMSs. Section 3 describes the research setting and Section 4 describes the variable measures and research sample. Section 5 reports the analysis results and Section 6 summarizes the results and discusses limitations of the study.

2. Review of Relevant Literature and Research Question

2.1 Theory of Feedback and Information Value

The decision-facilitating role of management accounting refers to the use of data as feedback to diagnose progress toward goals and as "feedforward" to devise interventions and course corrections that increase the likelihood of achieving those goals. Casas-Arce et al. (2017, 32) describe performance management systems as helping the "employee determine how best to allocate effort to improve decision-making" with reference to attention direction and problem solving. Sprinkle (2003, 302) identifies reduced ex-ante uncertainty and belief revision as the mechanisms through which feedback improves "…employees' knowledge, thereby enhancing their ability to make organizationally desirable judgments and decisions and better-informed action choices." Data-based decision making in education is defined by Ikemoto and Marsh (2007, 108) as "teachers, principals, and administrators systematically collecting and analyzing data to guide a range of decisions to help improve the success of students and schools."

Evidence that ASPIRE teachers use the PMSs to think deeply about remediation and to prioritize action for different learning objectives and targeted student intervention is plentiful in field interviews, as exemplified in remarks below:

> *"We use data a lot. I mean we're always using it to guide our instruction, to see where our students are weakest, what standards they're weakest in and trying to figure out what we need to do to try to fill those gaps. So we are using data all the time… [the PMSs] give me analysis breakdown based on the different standards and then I can go in and group students to figure out who's missing these certain areas. I can form small groups based on that information."*

> *"I can find out where kids lay out and it [the PMS graphical displays] color codes them… you've got your lowest*[2] *and then you've got your next kids, which we call our "cusp" kids. Those are the kids that are showing right now that they'll probably test on par. But with a little nudge, I can get them to proficiency. And those are the kids I really targeted for small group work. It could potentially raise their scores… for five weeks I met with them every Friday after school… I said "You are here because you're a good strong student, but we want to make you a super student, and you're right on the edge. And I showed them. I was very frank and showed them [their results]."*

> *"The [PMS] allows [me] to really break things apart into such small amounts of information about a student that you can really see where they're struggling…it might be reading comprehension holding them up in the writing area…you can really focus on reading comprehension and that's going to build performance in other areas."*

In addition to the substantive role of feedback, a PMS may also improve decision-making processes by making existing information easier to access without supplying "new" information. In research interviews, teachers indicated increased efficiencies in data retrieval. However, as these quotes

[2] In the interview, the teacher makes clear that the lowest performers are not neglected, but rather are typically special needs students who are included for a portion of class as part of mainstreaming. They are typically covered by individual education plans and receive individualized instruction from teachers outside of the main classroom.

highlight, this is confounded with spending more time with the data because it is easier to obtain and provides additional insights through new graphical presentation formats:

> *[The PMS provides] quicker access to a lot of information…if you're looking at data you can hover over something and see the number of students or the actual test question… [it] gives you access to looking at your data in all different ways, in a big picture."*
>
> *"You can get in there and interact with [the data] and it's not as time-consuming. You can see more in a shorter amount of time than in the past where we had to kind of search for it."*

Modern PMSs deliver information through a technology platform that integrates data from many sources. Access through a graphical user interface facilitates decentralized, frequent data use and investigation. The PMSs that we study supply teachers with daily feedback on student learning achievements and support trend analysis and other meaningful comparisons in both numeric and highly intuitive visual formats.[3] Their design and implementation are consistent with models of effective feedback and decision support (e.g., Keuning et al. 2017, Stecker et al. 2005). Teachers received focused training in using the systems and have a local "data driver" onsite to support their analyses. The dynamic data display, cues to guide analysis of the data, and suggested modes of intervention are all co-variates of success in a meta-analysis of experimental studies that examine the effects of data-based decision making on student performance (Staman et al. 2017 describing evidence in a Dutch publication by Faber and Visscher). These results echo Cardinaels and Van Veen-Dirks (2010) who used accounting data in an experimental study of the effect of display format on decision making, finding that graphical displays enhance the decision quality of users who are less proficient with numeric data. Many teachers emphasized the value of data visualization:

> *"…you can see from the bars [chart] how many kids answered each question and you get an idea that that's a test question that [was] really a challenge for everybody. I can tell at a glance…the strength [of the PMS] would be in laying out for me that stratification of kids that are at the very low end and the kids at the very high end… I've found that every year that I'm able to push kids up."*
>
> *"I can look at trend lines and say, "Oh, he was here. He was proficient in fifth-grade math. He went to basic and according to predictions for the interim [test] he is going to be below basic. What kind of intervention could we do to prevent that?"*

[3] The reporting and data visualization module is powered by Tableau, a data visualization software program that is widely used in conjunction with "big data" and business analytics applications.

The visual displays use color to highlight performance thresholds. Common reports include textual cues to remind the teacher what to look for and to suggest actions that may be warranted.

In sum, feedback has as its core an informed decision maker, and a complete test of the feedback effects of PMSs starts with a decision maker's use of information. Consistent with the "information value" of accounting data, PMSs are purported to deliver improved decisions and improved managerial understanding of real-world relationships. The mechanisms of improvement are similar to the "plan-do-check-act" cycle in operations management and processes of management control.

2.2 Evidence on effects of PMSs

For decades, the corporate sector has embraced performance management and the decision-facilitating role that PMSs play. Rigby and Bilodeau (2013) report that scorecards are one of the top ten management tools used by firms. Similar trends in the education sector are more recent but have spread quickly (e.g., Ikemoto and Marsh 2007, Staman et al. 2017). Management accounting research on whether feedback from PMSs is associated with superior organizational outcomes has produced mostly affirmative results, but they are not free of confounding effects. Evidence from the education sector is more limited and mixed, and while free of some confounding effects, suffers from other empirical challenges.

In the management accounting literature, experimental research provides conclusive, direct evidence that PMSs affect managers' decisions. Using methods from the judgment and decision-making literature, studies examine how the presentation, construction, and components of PMSs affect managers' performance evaluation and decision quality (e.g., Lipe and Salterio 2000, Vera-Munoz et al. 2007, Cardinaels and Van Veen-Dirks 2010, Cheng and Humphreys 2012), effort allocation (Farrell, Kadous, and Towry 2012), and the understanding of and ability to execute strategy (e.g., Taylor 2010, Humphreys and Trotman 2011, Cheng and Humphreys 2012). The evidence from these studies that PMSs influence users' understanding of their environments and shape their decisions is convincing. What this research cannot address is whether, in a natural setting, these effects deliver better organizational performance. Moreover, experiments make data and the problem salient; typically, participants receive performance data and are asked to render a decision or judgment (Sprinkle 2003). Thus, these studies are unavoidably silent about whether managers routinely consult PMSs in decision-making without prompting.

Empirical research situated in natural settings addresses some of these concerns, but has other limitations. For example, Ittner et al. (2003) examine the association between firms' financial outcomes and two facets of PMSs: whether firms use a diverse set of performance measures and whether the measures are aligned with strategy. They find that firms employing a more diverse set of metrics than competitors who pursue similar strategies have higher stock price returns but no higher accounting returns. When they consider direct, self-reported measures of whether the firm uses a scorecard or a causal business model, they find little evidence that these practices are associated with improved financial performance, but they are associated with greater user satisfaction. The authors acknowledge that endogeneity in adoption decisions (so called "selection effects"), ambiguity about what defines scorecard adoption for survey respondents, and model misspecification may explain the findings.

Davis and Albright (2004) address some of these concerns in a quasi-experimental study of a bank that implemented a balanced scorecard in one geographic segment but not another. They find that bank branches with a scorecard have higher subsequent performance compared to branches without one. Although branches are assigned to one treatment or the other, the firm's decision to adopt a scorecard is endogenous. Thus, the estimated returns to adoption may be biased. Moreover, the treatment is not fully randomized, so omitted factors may better explain performance differences than scorecard adoption.

These studies equate presence of a PMS with use of the PMS, and neither study isolates the feedback role of the PMS from incentives to use the PMS. Casas-Arce et al. (2017) addresses the linkage between the decision maker and organizational outcomes that is absent in earlier studies. They examine the effect of providing bank employees with data on customer lifetime value. Although incentive pay motivates employee behavior, the new information is introduced with no *change* in incentive pay. The authors conclude that the new information was used by employees to allocate effort to customers with greater potential lifetime value. However, consistent with the possibility that selection effects confound the feedback effects, managers with shorter job tenure exhibit a stronger response to the new information.

As compared with archival research, case study and survey researchers enjoy access to the decision makers who use PMSs and are able to provide direct evidence on the decision-facilitating processes that PMSs engender. For example, Malina and Selto's (2001, 75) case study finds that "managers respond positively to … measures by reorganizing their resources and activities, in some cases dramatically, to improve their performance on these measures." Similarly, in survey research Grafton, Lillis and Widener (2010, 689) find "…the extent to which decision-facilitating measures are actually used by strategic business managers impacts on the strategy capabilities of the organization *and subsequently its performance*." In both of these studies, incentives are confounded with PMS use for decision making. In addition, self-reported PMS use and self-reported performance give rise to concerns about reporting bias and measurement error.

In the education sector it is uncommon for PMSs to confound decision-facilitating and decision-influencing roles of the PMS because the sector has typically not made extensive use of teacher incentives. Government efforts to mandate performance pay for teachers has met with resistance. Moreover, Fryer's (2013) large-scale randomized field experiment in over 200 New York City public schools provides persuasive evidence that there is no causal relation between incentive pay for teachers and student learning outcomes, attendance, behavior, or graduation rates. Firestone and Pennell (1993) posit that teachers' motivation to improve student performance derives from personal commitment and the intrinsic rewards of student achievement rather than a desire for contingent pay. Thus, the sector provides an opportunity to isolate the decision-facilitating effects of PMSs.

In spite of policies that encourage or require teachers and schools to use data to improve overall performance, education research on the efficacy of PMSs is limited (Staman et al. 2017). Often the research is conducted with small groups of students and teachers, over a short duration, and without an experimental design to address self-selection by students or teachers to be study participants. Like studies of PMS efficacy in the corporate sector, few studies link *actual use* of PMS systems by individual decision makers with organizational outcomes. For example, Ikemoto and March (2007) rely upon educators' self-reported use of data from interviews to assess whether PMS use is associated with certain educator actions but not with student learning outcomes. Slavin, Cheung, Holmes, Madden and Chamberlain (2013) use an experimental design in which some schools use a particular data-driven reform while control schools do not to compare average student achievement at various grade levels in English and math. They find evidence of positive effects of the presence of a PMS in some grades for some subjects; however, importantly their data are school-level and do not measure *actual use* of the PMS by teachers. Staman et al. (2017) use experimental design with treatment (42) and control (42) schools, with staff in treatment schools receiving access to and training in how to use interim student assessment data to identify improvement opportunities. They find no main effect on student-level mathematics outcomes of being in a treatment school but do find an interaction effect between treatment and the student having low prior achievement and low socioeconomic status.

These studies illustrate how the research question has been examined and offer some of the best extant evidence that PMSs deliver improved organizational outcomes. Nonetheless, they also illustrate the difficulties of establishing the causal chain of 1) a decision maker's use of the PMS, 2) actions by the decision maker that are informed by the PMS, and 3) improved organizational performance. Confounding effects of incentives and of selection have made it difficult to assess effects of feedback on performance.

2.3 The Research Question

This study provides evidence on the feedback role of performance information. The research question, stated in terms of the specific research setting, is very simple: are teachers who make greater use of PMSs associated with better student learning outcomes? Like the thousands of studies summarized by Hattie (2009), the maintained hypothesis is that the focal education intervention (PMS use) is associated with material positive effects for learning. Although the question is straightforward, its answer is neither trivial nor assured. The research site adopted PMSs voluntarily with the expectation of positive effects on student learning and teachers voluntarily use the PMSs.[4] Thus it may seem that the hypothesized effect is

[4] Institutional theory (Meyer and Rowan 1977, DiMaggio and Powell 1983) explains the phenomenon of an organization adopting a technology but *not* using it. Adoption does *not* lead to performance improvement either because of disuse or perfunctory, symbolic use. This theory does not provide a compelling counterargument for the research hypothesis of this study because we find no evidence that teachers experienced pressure to use the PMS.

assured. However, even well-designed PMSs may fail to yield material improvements in learning outcomes if 1) PMS use comes at the expense of more effective teaching activities, 2) PMS use produces results that are weakly related to learning but support other valued outcomes (e.g., prosocial behavior), or 3) the PMSs confirm what the teacher already knows, providing no incremental information (Keuning et al. 2017). Field interviews indicate that these threats are present at the research site.

Teachers operate with tight time and resource constraints and are charged with achieving a variety of outcomes. Staman et al. (2017, 59) note, "Changing teacher behavior is not easy, because of the many obligations [that they] face in their work." Evidence suggests that the primary management challenge of schools is teacher effort allocation rather than teacher motivation (Fryer 2013, Firestone and Pennell 1993). Several teachers mentioned these impediments to using the PMSs in interviews:

> *"…it was not a lack of wanting to look at data. It was the time involved in collecting it. [Teachers] … are teaching and stuff happens… all the time spent on just collecting or getting your data ready rather than looking at it and responding… it took so much energy to just try to get the data… [teachers] have very limited time. They have tomorrow to plan for and they have a stack of papers to grade…"*
>
> *"To be honest with you, it is not the [PMS] tools' fault. It is our bandwidth…"*
>
> *"I don't feel like I'm a particularly proficient user…it would just take time slogging around but we have so much on our plates…"*

Time teachers spend analyzing and reflecting upon the data in PMSs could be spent with students, parents, in lesson planning, or working with other teachers. For PMS use to positively influence student outcomes, it is not enough for it to be effective in a vacuum (e.g., an experimental setting). It must have a higher marginal return to important outcomes than alternative uses of teacher time. Keuning et al.'s (2017, 32) case studies of "successful" and "unsuccessful" uses of data-based decision making in schools noted that teachers in failing schools perceived the PMS as "time consuming and involving a lot of paperwork." If data analysis using the PMS consumes a great deal of teacher time at the expense of more effective ways that teachers influence student learning, the research hypothesis will be rejected.

We examine the association between teachers' PMS use and student's academic outcomes; however, teachers are also concerned about students' pro-social behaviors, including attendance and engagement in the learning community, because they are charged with educating the whole person and developing the next generation of citizens. Several field interviews included remarks from teachers about using the PMS to communicate with students or parents and to collaborate with other teachers. An earlier quote described how one teacher shared visual data from the PMS with students who were on the cusp of proficiency to motivate them. Two other teachers described the value of the PMSs in parent conferences:

> *“It is nice when I’m reviewing report cards and, even in student-led parent conferences, when we can just click on the kid’s name and pull up all of their stuff. It’s a time saver too, because if I’m having a conversation with a parent I can easily pull up all of their data and say “look how they did!”*

> *“I really use data for communicating to parents, that’s top priority. We have what’s called a student-led conference twice a year and that is where I pull data from on where they need to grow. Then I show the parents, ‘Oh, your student had this percentage in this subject this fall and now in winter, they had this.’ Talking about growth and what I need to see with them for growth. It is also very helpful with my report cards to be able to show growth or no growth.”*

The PMS allows teachers to compare performance of their students with those of other teachers who teach the same course. Several teachers described how the PMS promoted collaboration and sharing teaching materials. Parent and student communication and teacher collaboration are mechanisms for supporting student learning; however, if PMS use is associated with these purposes with no subsequent impact on learning, the hypothesis of positive association will be rejected. We do not have data to permit a direct test whether these intermediate outcomes are associated with PMS use.

The premise of PMSs is that they inform; they reveal something that the decision maker did not know or they reduce uncertainty about what the decision maker believes to be true. Education research clearly indicates that, after controlling for student demographics and family influences, “teacher fixed effects” explain the largest portion of variance in student learning outcomes (e.g., Hattie 2009), yet we know comparatively little about the systemic differences among teachers that create these effects. Evidence suggests that the differences are unrelated to readily observable characteristics (e.g., education, experience) and are more likely to be found in classroom behaviors and teaching practices (Goldhaber 2002, Hanushek 2006, Hattie 2009, Stronge et al. 2011). We hypothesize that using PMS data to identify remediation opportunities and structure lesson plans accordingly is one such practice. However, for teachers' PMS use to influence student learning, the PMS must provide information that is incremental to what the teacher already knows or it must help the teacher arrive at that knowledge more quickly. Although it was rare, perhaps due to a bias toward agreeable responses, some teachers raised this point. For example, when asked about the information that one of the PMSs offered, one teacher replied. "Nothing. I mean, you want honesty, I'll give you honesty. We are so overwhelmed with data… I stick with what I've used for the last ten years. I know this stuff." In Keuning et al.'s (2017, 32) unsuccessful school, teachers complained that the PMS "failed to support their education practice." Teachers use formal and informal means for assessing student learning. If PMSs are used to confirm what they already know but do not add to their existing knowledge, then the research hypothesis will be rejected.

The above arguments identify threats to the hypothesized positive association between PMS use and learning outcomes. As earlier studies indicate, an equally important threat relates to interpretation of any documented statistical association. In a natural setting, decision makers' choice to use the PMS is endogenous and likely related to characteristics of the decision maker. If "better" teachers use PMSs, support of the research hypothesis requires an identification strategy to empirically disentangle the hypothesized feedback effect from selection effects. Evidence of this threat is found in Dunn, Airola, Lo and Garrison's (2013) study modeling teachers' decisions to use data as a function of perceptions of efficacy in using data and anxiety about data use. We employ identification strategies to distinguish feedback effects on PMS use from selection effects.

3. Research Setting

3.1 The Research Setting: The Education Sector and performance measurement

Economists, legislators, and business leaders worldwide have reached uncommon unanimity on the importance of education to economic growth. While the preferred tactics for improvement differ, common themes around accountability and assessment have brought performance measurement systems to the fore. In the U.S., impetus for these solutions comes from federal funding for states that is tied to accountability and assessment standards (e.g., "No Child Left Behind (NCLB)" and "Race to the Top" programs) and from non-governmental organizations that sponsor education reform (e.g., Gates

Foundation, Dell Foundation, Spencer Foundation, Broad Foundation). An influential international study by McKinsey & Co. (2010) of 20 exemplary school systems that implemented 575 distinct reforms underscored the importance of assessment practices to improved school performance.[5]

Standardized tests have been the backbone of assessment of student learning, commanding a disproportionate share of attention (Stiggins 2005).[6] With performance dashboards, teachers have turned from assessment *of* learning to assessment *for* learning (Bennett 2011). Like dashboards and scorecards of the private sector (Kaplan and Norton 2004; Rigby and Bilodeau 2013), new assessment practices in the education sector feature: data visualization, a focus on time trends and cross-sectional comparisons, multiple measures of performance including leading indicators of test performance (e.g., absenteeism, timely submission of homework), and "drill down" capability to facilitate different units of analysis (e.g., student, class, teacher, subject, school).[7] Assessment feedback helps teachers identify and remedy student learning difficulties before high stakes tests are administered (Black and Wiliam 1998) and are implicated in school success (McKinsey & Co. 2010; Childress, Elmore, Grossman and Johnson 2007).

Founded in 1998, ASPIRE Public Schools is a large charter school[8] organization that, in the 3 school years of this study, educated approximately 14,600, predominantly low-income, K-12 students in 38 schools, primarily in California. The school's motto, "College for Certain," speaks to its mission of preparing students in underserved communities for college. ASPIRE fulfilled federal requirements to test students regularly in English language arts (ELA) and mathematics using the California Standards Tests

[5] The study finds that establishing reliable data systems and conducting regular, standardized assessments of student learning are critical to sustained performance improvement.

[6] While tests are unlikely to measure all desired educational outcomes, Chetty, Friedman, and Rockoff (2014b) provides large scale, longitudinal evidence that value-added measures based on improved test scores are leading indicators of outcomes such as college attendance, earnings, and a variety of long-term pro-social behaviors.

[7] Rethinam (2014) describes the emergent uses of data analytics and predictive modeling in the K-12 sector. Kaplan and Lee (2007) provide a case study of the implementation of a balanced scorecard in a public school district.

[8] A charter school is a public school that students choose to attend at no cost to their family. ASPIRE imposes no admission requirements; however, if demand exceeds capacity (a rare occurrence), enrollment is determined through a lottery. Charter schools are granted their charter from a state and receive state support for each student enrolled, but are freed from some state regulations (e.g., operating hours, hiring practices). Like many charters, ASPIRE uses smaller schools, smaller class sizes, extra school days, extended school hours and a modified school calendar with shorter summer recess than its public school counterparts.

(CSTs), to evaluate student progress toward meeting academic standards for each grade and subject.[9] CSTs are measured on a scale from 150 (low performance) to 600 (high performance). For the students in our analysis, in Year (2011-12, 2012-13, 2013-14) the mean [median] ELA CST was (367 [367], 368 [367], 364 [364]) and math CST was (399 [394], 392 [388], 390 [385]). The distributions of scores, (Figure 1) indicates that although the mean and median are centered, there are important differences in ELA and math. For both subjects the lower tail is 'heavy' with low-performing students (kurtosis by year for math [ELA]: (3.03 [3.19], 2.99 [3.45], 3.07 [3.57])). The ELA distribution is approximately normal, but the math distribution is skewed by a significant number of high-performing students. The data are consistent with ASPIRE's focus on "at risk" youth, many of whom are not native English speakers.

3.2 The Decision-facilitating Uses of Two Performance Measurement Systems (PMSs)

ASPIRE has a strong data-driven culture. Proximity to Silicon Valley and success obtaining competitive grants have shaped its use of technology to meet accountability requirements. We conducted field interviews with 17 teachers and principals from five schools in Spring 2012 to understand PMS use before obtaining access to archival data. ASPIRE teachers use two online, interactive, dashboard-style PMSs called the data portal, or "portal", and "Edusoft®" to analyze current and historical student performance, to identify problem areas, and to devise lesson plans for improvement.

The portal is an interactive, data visualization interface that was designed in-house. It allows teachers and principals to analyze student progress in relation to statewide grade-specific learning objectives and to create customized reports at the student, class, and school level.[10] Appendix Figure A-1 provides two examples of displays that a teacher might use to identify students who are in need of remedial work on specific learning objectives. Cues prompt the teacher about how to interpret and use the

[9] Details about the California Standardized Testing and Reporting Program (STAR) are at: www.startest.org . The CSTs were discontinued in 2014 in conjunction with adoption of the Common Core State Standards.

[10] See http://schoolzilla.org for additional information about and examples of this dashboard system. The ASPIRE group that developed the data management platform in 2009 received grants to expand the capabilities of the program and was spun off to form an independent social enterprise in 2013. The new organization, Schoolzilla, provides free and fee-for-service products to schools nationwide (http://blogs.kqed.org/mindshift/2012/09/charter-school-network-offers-its-own-data-system-to-all-schools/#more-23757).

data. The teacher can also compare aggregate class performance on specific learning objectives to other teachers' classes. The portal supports benchmarking and helps teachers identify other teachers with whom to collaborate. Importantly, the data in the portal are not new to the organization or the teacher. Rather, the portal makes data readily accessible and useful by allowing teachers to search and filter data on a variety of student, subject, and school characteristics and for different time periods, and by converting data into visual displays. It draws from a common data warehouse that includes item-level, standardized test performance as well as student attendance and demographic data.

The portal was launched and all teachers and principals, were trained in its use early in 2010. In each school a designated 'data driver' (also a classroom teacher) helps teachers use data to answer questions, conducts special studies at the request of the principal, and supplements teacher training. Headquarters staff members also provide support, analyzing student performance in support of grant applications and state accountability reporting. Teachers who report heavy use of the data portal PMS emphasize the importance of assessing interim performance to identify targets for remediation before the end-of-year CST is administered. Those who report light use of this PMS tend to dismiss the diagnostic value of intermediate tests or to indicate personal difficulties using the PMS or interpreting its reports.

A second PMS, Edusoft®, is a third-party, web-based assessment platform that has a long history of use by ASPIRE teachers.[11] Edusoft® is a full suite of curriculum management and student assessment products; however, ASPIRE teachers use it primarily to link detailed learning objectives and state grade-level standards with student-level assessments of mastery. The system is best understood through an example. A middle school math teacher described how he and his fellow teachers developed weekly assessments using a combination of teacher-generated problems and problems taken from the Edusoft® Item Bank. Each problem, regardless of source, is coded to at least one state standard. The Edusoft® platform is used to print the test answer sheets and students complete the multiple-choice questions by

[11] See http://www.edusoft.com/corporate/about.html for a more complete description of the technology. Several teachers who we interviewed indicated that their use of Edusoft predated joining ASPIRE because it was also used by their prior employer, the local public school district.

‘bubbling in’ their response. (The test allows for open-ended or essay questions; however, for open-ended questions the teacher records grades onto the test sheet using a similar ‘bubble’ marking. After the open-ended questions are graded, the teacher scans the sheets using an office scanner connected to the internet.) Edusoft® grades the assignment and records the student’s score for each test item in the teacher’s electronic grade book. Because test items are cross-referenced to learning standards, teachers can manipulate the data to identify commonly missed items and learning standards that should be re-taught to all or a subset of students. The power of the PMS is limited only by the degree to which teachers are able to use its testing materials to assess classroom learning. In interviews, teachers who are most enthusiastic about it highlight the detailed diagnostics for individual learning objectives. Teachers who report making less use of Edusoft® offer such explanations as: very young students are unable to reliably use the ‘bubble’ test answer forms; the Edusoft® test bank is incomplete; and, ELA standards are not as amenable to evaluation with the multiple-choice format as Math standards. Differences in the perceived usefulness of PMSs for math and ELA instruction foreshadow our analysis and shed light on prior research that finds that teachers and charter schools have greater success lifting student math performance than ELA performance (Hanushek and Rivkin 2010; Tuttle et al. 2013).[12]

The two PMSs are separate but complementary diagnostic resources. Edusoft® helps principals and school administrators demonstrate that the school curriculum meets state-level required standards and supports teachers in assessing student mastery of these standards throughout the school year. However, the information in Edusoft® does not become a part of the student’s permanent historical record and teachers can only access data for their own courses and the students enrolled therein. The portal is a comprehensive, holistic profile of the student that covers all of their years in the ASPIRE school system.[13]

[12] Importantly the regular finding that math scores are more malleable than ELA scores does not translate into lower importance for later outcomes. Chetty et al. (2014a) find that teachers’ ELA value-added matters as much or more to long term student outcomes than Math value-added.

[13] At present, teachers are permitted complete access to individual student records for any student currently enrolled in their courses and may access aggregate classroom statistics for other teachers and for other schools. Thus for example, an 8th grade Algebra teacher may compare the performance of her class with that of all other 8th grade Algebra classes in the same school, or in any other ASPIRE school. ASPIRE provides transparency to encourage relevant comparisons and learning within and across schools, while protecting student privacy.

Its performance focus is the required annual end-of-year state tests (the CST). The logical connection between the two PMSs is that student mastery of the state standards is expected to be a leading indicator of performance on the CST. From field interviews, it is clear that the PMSs are not substitutes for one another; the portal and Edusoft® are used to varying degrees depending on perceived 'fit' with the teacher's assessment style and the subject matter. To a person, the teachers and principals interviewed described ASPIRE's 'data driven' culture. Placards and bulletin boards in the schools attest to this, with visual presentations of student and school achievement in every corridor. Although teachers are neither compelled nor incentivized to use the PMSs, teachers report using them as a means to the desired end.

4. Variable Measurement and the Research Sample

4.1 Dependent Variable: Student-level Performance Improvement

Federal accountability requirements focus on annual student *improvement* ("growth") rather than attainment of a specific achievement *level*. This recognizes that a student's absolute achievement level may be influenced by many factors outside of the school's (and teacher's) control. ASPIRE measures 'student growth percentiles' (SGP) (Betebenner 2008). SGP measures how a student's *improvement* in CST scores between two consecutive school years compares to improvement of non-ASPIRE students in the same grade who earned the same CST score in the prior year. ASPIRE uses all students in the Los Angeles Joint Unified School District as the population of peer students against which its students are compared. Comparing students with the same 'starting point' addresses concerns about mean reversion tendencies that influence raw test scores. SGP scores range from 1 to 99, reflecting the percentile of the student's *ex post* performance within the distribution of peers with a common *ex ante* CST performance. In sum, while the CST measures *absolute* grade-level-specific subject mastery, the SGP measures *growth* in the CST in relation to peers with the same starting level of subject mastery.[14] The dependent variables of this study are student SGP scores for math, *Mathsgp*, and ELA, *ELAsgp*, in three school years.

[14] In the population of peer students, start-of-year CST and end-of-year SGP are uncorrelated by design (i.e., for each initial level of CST, the distribution of final CST scores is translated to a normal distribution and the percentiles of this distribution, SGP, range from 0 to 100). For the population of ASPIRE students, it is possible theoretically to obtain a nonzero correlation between CST and SGP if ASPIRE students collectively exhibit

SGP is used in reports for federal and state accountability; but during the period of study, it was not widely shared with teachers. In interviews we found no teacher who knew the SGP of his or her students, a few who were familiar with the concept and had been told by their principal the average student SGP of the school, and many who were unaware of SGP. In contrast, all teachers were keenly aware of their students' performance on the CST. During the period of study, neither SGP nor CST was a component of teacher compensation; that is, there was no contingent 'pay for performance' for teachers.[15]

4.2 Independent Variables: Teacher Portal and Edusoft® Use

We use system records of teacher logins and downloads to measure the frequency of use of the two PMSs. A strength of this approach is that it is an unobtrusive measure of actual use rather than a perceptual, retrospective, self-reported measure of use. Prior to our request for the records and creation of these metrics these usage measures did not exist. PMS use was neither actively monitored nor reported to teachers or principals. Since the measures reflect teachers' personal usage choices (e.g., PMS use is not a randomized treatment) we are mindful of self-selection in PMS use.

The login and download data have some limitations. For the portal, the system only records the date and time that a teacher accesses a report. The primary measure, *portal use*, is the total number of reports downloaded by a teacher during the school year.[16] Because Edusoft® is a third-party, web-based tool, ASPIRE does not have direct access to daily usage records. We requested and received a file from the vendor containing data on the cumulative logins to Edusoft® for each teacher for each school year. Comparing teachers' use and nonuse of both PMSs, in the three school years, 83, 70, 83 % of teachers

systematic under- or over-performance. The correlation is very low for the school years that we study: in year 1 (2, 3) the correlation between CST and SGP for math is -0.11 (-0.01, -0.06); for ELA it is -0.05 (-0.02, -0.06).

[15] ASPIRE schools are located in six large California districts (termed 'CORE') that in 2013 received waivers from NCLB. NCLB requires standardized tests to be used in teacher evaluations. The Governor opposed test-based teacher accountability, so to obtain a waiver, the CORE districts agreed to work toward school evaluation systems that include student learning measures (The districts' request for waiver is available at: http://www2.ed.gov/nclb/freedom/local/flexibility/waiverletters2009/cacoreflexrequest22013.pdf). The waiver was granted and has since been extended, with evidence of progress toward piloting new school evaluation systems in 2014-15 and teacher evaluation systems in 2015-16 (Fensterwald, 2015a, b).

[16] In untabulated analysis we considered as an alternative measure the number of days in the school year for which there is at least one login (to discriminate a prevalent pattern of use from a user who downloads a large number of reports in a single day). The measures were highly correlated and produced similar results to those that we report, indicating that for most teachers, usage is distributed across the school year.

made some use of both PMSs, 3, 18, 6 % used only Edusoft®, 11, 9, 10 % used only the portal and 4, 2, 1 % used neither PMS. Clearly, the norm is to use both PMSs. For each student we develop a student and subject-specific, 'teacher PMS usage' measure that is the average PMS usage by the student's teacher(s) (*S-Portal use, S-Edusoft use*). Students are taught a subject by up to four teachers.

4.3 Control Variables

We use several methods (described later) to address selection bias associated with better teachers making heavier use of PMSs, but as a starting point we include as a control variable an independent measure of teacher effectiveness. Starting in 2012-2013, the second year of study data, and each year thereafter, teachers are evaluated by their school principal using a 100-point scale. Among teachers who worked at least one year of the three in this study, 65% were evaluated at least once.[17] We measure *teacher effectiveness,* as the average of the principal ratings that the teacher received in 2012-2013 and 2013-2014 (using a single year rating if there is only one). For each student we develop a student and subject-specific, average teacher effectiveness measure (*avg teacher effectiveness*) by averaging the ratings of teachers (up to four) who taught the student the specific subject.

In education research, education level and years of teaching experience are common proxies for teacher effectiveness . All of the teachers in the sample have a bachelors degree and only one has a masters degree. Personnel records include sparse data on teachers' years with ASPIRE and prior teaching experience. Prior research is ambivalent about the merit of these controls. Hanushek (2006, p.1061, Table 1) summarizes 34 high quality studies. None finds a positive significant association with a teacher having a masters degree and only 41 percent find a positive association with experience. Rivkin, Hanushek and Kain (2005) conclude that any positive effect of experience is better described as a negative effect of the first year of teaching with "essentially flat impacts of experience subsequently." These studies do not

[17] We lose 274 observations associated with 132 unique teachers who have no principal rating. Of these, 115 teachers were not working in 2012-14 (i.e., they were employed in earlier years before ratings were given), eight joined Aspire in 2012-14, and nine have no rating for reasons that we cannot determine. In the results that follow we estimate the models with and without the teacher effectiveness variable to ensure that the presence of a rating is not itself a source of selection bias that alters the results.

include a direct measure of teacher effectiveness. For teachers in our sample with both employment data and principal effectiveness ratings, the correlation with years teaching for ASPIRE is 0.16 (p=0.03) and for total years teaching is 0.15 (p= 0.05). In untabulated results that include both the teacher ability rating and years of experience (with a reduced sample), both variables are positively associated with student growth; however, neither the magnitude nor the significance of the hypothesized associations are affected. The second measure of ability (experience) reduces the magnitude but not the significance of the principal rating with no influence on the association between PMSs and student growth, suggesting that principals factor experience into their own ratings. The use of a difference-in-difference approach in which the teacher serves as her own control further mitigates the need for additional teacher-level controls that might be considered if the data were available.

Student demographics are well-established correlates of education outcomes and are commonly included as control variables. However, inclusion of these variables has been criticized as suggesting that achievement and growth are persistently impaired in a subgroup of students. The SGP approach of benchmarking growth among students with the same starting point purports to eliminate the need for controls. The demographics of the Los Angeles Independent School District that provides the peer group comparison for ASPIRE students are comparable and this should diminish the need for demographic controls. Notwithstanding these arguments, Ehlert, Koedel, Parsons and Podursky (2012) demonstrate that SGP often remains correlated with student demographics; thus, we include two controls for student demographics. The first is an indicator variable for students who are designated native English speakers (*English only*). This group excludes students who speak English as their second language (ESL) with varying proficiency. Instruction is in English, so one might expect native English speakers to have a language-based advantage that is perhaps more pronounced in ELA than in math. However, prior studies show that growth of ESL students is systematically under-predicted by models of 'normal growth' (Lakin and Young 2013). The latter finding means that ESL status is associated with *greater* growth than *English only* status. We make no predictions about the sign of this control variable, in consideration of divergent explanations. The second control is an indicator variable for students from an economically disadvantaged household (*Disadvantaged*), an effect that is expected to impair learning growth (Henry, Thompson, Fortner, Zulli and Kershaw 2010).

4.4 Research Sample and Data Preparation

The research sample includes all ASPIRE students in grades 3-8 for three school years. Assessment of SGP requires an initial and a subsequent CST score and the CST is first administered in grade 2, so grade 3 is the first year for which SGP is established.[18] The population of 7,537 students in grades 3-8 is reduced by 923 students who are classified as 'disabled.' Although these students take the CST, federal rules allow their progress to be monitored separately. In addition, we eliminate 99 observations for students who cannot be linked to a teacher who teaches 5 or more students. The resulting research sample has 6,515 unique students who are associated with 12,687 student-years of data (12,618 math SGP, 12,649 ELA SGP, and 12,580 in which both subjects are present). Over the three years, 383 unique teachers taught these students; 314 taught math and 309 taught ELA (in lower grades, teachers typically teach both subjects while in upper grades teachers specialize). Four or fewer teachers teach students a subject in one school year. The student-teacher records are associated with one of 30 ASPIRE schools in full scale operation during the year, each with its own principal and leadership team.

The data have a multi-level, but not fully nested structure: teachers are associated on average with 28 students[19] and students are associated with up to four teachers. Because students are not fully nested within teachers, we have two options for linking teacher behaviors (PMS use) with student outcomes (learning growth). We can disaggregate teacher behavior to the students they teach, or we can aggregate student learning outcomes for all students that a teacher teaches. The approaches provide complementary evidence; however, for parsimony in presentation we focus on the student-level tests, which require the least amount of data aggregation, and use teacher-level tests for robustness checks.

4.5 Variable Descriptive Statistics

[18] Although the CST is administered for higher grades, the tests are more differentiated (e.g., Algebra I, II, Geometry). In the relatively young ASPIRE school system, which started with lower grades and grew as the students aged and high schools were added, the sample size (students and teachers) for each test would be quite small.

[19] The data do not include information on 'classes,' groups of students who learn from a teacher in one time period.

It is useful to consider whether it is necessary to estimate separate models for the effects of PMS use on math and ELA SGP. Figure 1 provides preliminary evidence on how ELA and math CST scores differ. For student-years with both a math and an ELA SGP score, the correlation between these measures is 0.28. Thus, although general ability may create positive correlation between ELA and Math achievement, there is considerable unique variation to warrant separate analysis.

Table 1, Panel A provides variable descriptive statistics by subject. Both mean *Mathsgp* and mean *ELAsgp* are 50.5, indicating that ASPIRE students are well-matched to the Los Angeles peer group, and the full range of SGP (1 - 99) is observed. Descriptive statistics for use of the PMSs are reported at the student-level, with students who share the same teacher(s) having the same entries for the PMSs. (Appendix A-1 provides descriptive statistics at the teacher level.) For the portal, teachers' mean annual use is approximately 31 downloaded reports for both subjects. For Edusoft® it is 161 logins for ELA teachers and 170 for Math teachers. For both PMSs a wide range of use is observed from no use (0 reports, 0 logins) to high levels of use (269 reports, 753 logins). Importantly, variation in PMS use indicates that even if teachers self-select into the "data driven" ASPIRE organization, they do not feel compelled to use the PMSs. Eighty percent of students are from disadvantaged households and 48 percent are native English speakers. The average effectiveness rating of the 292 unique teacher combinations for whom ratings are available is 87.4 with a range from 21 to 100 and a standard deviation of 15.6.

Panel B tabulates univariate variable correlations separately for each subject. Consistent with the hypothesis, *ELAsgp* exhibits weak but significant positive correlation and *Mathsgp* exhibits stronger significant positive correlation with both *S-portal Use* and *S-Edusoft Use*. The difference between the two subjects suggests that math may be more amenable than ELA to diagnostic assessment. As expected, t*eacher effectiveness* has a positive significant correlation with both *ELAsgp* and *Mathsgp*, but in both cases the relation is relatively weak. Modest correlations between SGP and student demographics variables suggest that the construction of SGP is somewhat successful in diminishing associations between performance and demographics. Several correlations among the independent variables and controls are statistically significant but the magnitudes are small enough to mitigate multicollinearity. The two PMS use measures are positively correlated at the 0.18 level; however, because we are interested in the joint effect of PMS use rather than the precise effect of either PMS, multicollinearity is not a major concern. *Teacher effectiveness* is intended to control for better teachers making greater use of the PMSs. The univariate correlations indicate a positive significant association between *teacher effectiveness* and *S-portal use* (0.12) and a small but significant negative association with *S-Edusoft use* (-0.04).

5. Results

5.1 The Influence of PMS Use on Student Growth

To assess whether student growth is associated with PMS use by the student's teacher(s), we estimate separately for ELA and math SGP, the OLS regression:

Student [math or ELA] SGP = $\beta_0 + \beta_1$ (*S-Portal Use*) + β_2 (*S-Edusoft Use*) + $\sum_{i=1}^{n} \beta i \; (Control\ i)$ (1)

where the control variables indicate whether the student is *Disadvantaged* or speaks *English only*. We report models with and without the control for *average effectiveness* of the student's teacher(s).[20] Each unique combination of teachers that teach any student is treated as a separate cluster for computing robust standard errors and determining the significance of the coefficients. We include but do not tabulate (for purposes of confidentiality) a significant model constant and jointly significant school fixed effects.

The results, tabulated in Table 2, indicate that *ELAsgp* is positively associated *S-Edusoft use* ($p \leq 0.01$) even after controlling for teacher effectiveness. *S-Portal use* is not significant after controlling for teacher effectiveness. Evaluated at the mean value of *S-Edusoft use* (161.3* 0.013), the estimated coefficient indicates that teachers who make average use of the PMS raise student *ELAsgp* 2.1 percentiles compared with those who do not. *ELAsgp* is negatively associated with student demographics, with *English only* students scoring 1.0 percentile less[21] and *Disadvantaged* students scoring 2.0 percentiles less, on average. Thus, a teacher who makes average use of the Edusoft PMS offsets the negative association with ELA growth of student household poverty. *Mathsgp* is positively associated with the use

[20] We include both models to demonstrate that any selection bias associated with eliminating students taught by teachers without a teacher rating does not affect results.

[21] Lakin and Young (2013) find that 'normal growth' models under-predict ESL student growth.

of both diagnostic tools ($p \leq 0.01$ for *S-Portal Use* and $p \leq 0.01$ for *S-Edusoft Use*) after controlling for teacher effectiveness. The differential responsiveness of *Mathsgp* and *ELAsgp* to PMS use mirrors findings in the education literature that math achievement is typically more responsive to all education interventions (Hanushek and Rivkin 2010; Tuttle et al. 2013). Teachers who make average use of the portal raise student *Mathsgp* 1.56 percentile points. Teachers who make average use of Edusoft® raise student *Mathsgp* 2.88 percentiles. On average, *Disadvantaged* status costs students 2.87 percentiles and *English only* students suffer a 2.2 percentile reduction in *Mathsgp*. Teachers who make average use of both tools almost offset the estimated negative association with math growth of student demographics (increase of 4.44 versus decrease of 5.07).

The estimated coefficients for the PMSs decline slightly with the inclusion of *teacher effectiveness*, consistent with selection effects. Consistent with other studies of teacher effects on student learning (Hanushek and Rivkin 2010; Tuttle et al. 2013), model fit statistics indicates that variation in math growth is somewhat better explained by education interventions than is the variation in ELA growth. A variety of explanations have been offered for this discrepancy. One that echoes our interviews with ASPIRE teachers is that the linkage between intermediate assessments found in PMS data and the final CST tests is stronger for math than for ELA. If true, then teachers who use the PMS will be better able to pinpoint problems and devise remediation strategies in math than in ELA.

To facilitate comparison of the results with other studies, we calculate the effect sizes (Grissom and Kim 2012) on student learning for teachers' use of the PMSs and for the students' exposure to teachers who are in the upper versus lower range of teacher effectiveness. We are unable to compute true effect sizes because there are few teachers who use neither PMS. However, we consider three "treatment effects" defined by above-median usage of the portal and Edusoft (BOTH), above-median usage of portal coupled with below-median usage of Edusoft (PORTAL), and below-median usage of the portal coupled with above-median usage of Edusoft (EDUSOFT), all as compared with below-median use of both PMSs and with inclusion of all control variables. Using Cohen's *d*[22], the measure that Hattie (2009) advocates for evaluating impact on education outcomes, the effect sizes of all three treatments are significant for math at the 95% level (Cohen's *d* is .27, .19, .08, for BOTH, EDUSOFT and PORTAL, respectively). For ELA, the BOTH treatment and the EDUSOFT treatment are significant at the 95% level (Cohen's *d* is .15, .06, .02). Separately, we examine the effect size on SGP of being taught by teachers with an above-median average teacher effectiveness rating, relative to a below-median average teacher effectiveness rating. Cohen's *d* = 0.14 and is statistically significant for both ELA and for math.

5.2 Evidence on targeted use of PMSs to achieve strategic objectives

ASPIRE teaches a high proportion of at-risk students with a central tenet of "college readiness for all." Investments in PMSs and training for teachers have been justified in part by the premise that early detection of learning difficulties allows teachers to develop targeted lesson plans for those in need of remediation. Asked to describe how they use the PMSs, most teachers described routines for clustering students with common learning difficulties to re-teach material. These factors suggest that the benefits of PMS use may be concentrated among students with the lowest realized growth. If so, linear regression over the full sample of students may be a poor tool for detecting differential effects of PMS use.

We employ simultaneous quantile regression[23] to assess whether the marginal effects of PMS use differ across the distribution of realized SGP outcomes (Koenker and Hallock, 2001). Table 3 compares the estimated coefficients for the two PMSs for deciles of the SGP distribution. Figure 2 is a graphical presentation of how the estimated coefficients from the quantile regressions differ from the average coefficient of the OLS regression. For each variable, the horizontal dashed line represents the OLS estimated coefficient and the two dotted lines on either side are the 90 confidence intervals for the least

[22] Cohen's *d* is calculated as: d = t [na+nb]/na*nb]^(1/2), where t is the test statistic associated with the treatment variable and na and nb are the respective sample size of the treatment and control group, respectively.

[23] In Stata, the *sqreg* procedure fits quantile regression models (also known as least-absolute value models) using bootstrapping and the variance-covariance matrix of the estimators to generate confidence intervals that facilitate comparison of coefficients between the quantiles.

squares estimate. The solid line connects the point estimates of the coefficient from quantile regressions and the shaded area around it is a 90 percent confidence band for these estimates.

Considering first *ELAsgp*, Table 3 and Figure 2a indicate that for all deciles the estimated coefficients on both PMSs use fall within the confidence interval of the OLS estimate. The confidence intervals for the quantile regressions suggest greater benefits of both PMSs in the lower half of the deciles; however, the results do not reach conventional levels of significance to support a conclusion of differing marginal effects for at risk students. Turning to *mathsgp*, Table 3 indicates that *S-Portal Use* has a significant positive effect for all deciles; however the effect is pronounced in the lower deciles and declines almost monotonically from the fourth to the ninth deciles. Figure 2c illustrates the increased efficacy of portal use in the lower deciles. For *S-Edusoft Use*, Figure 2d indicates consistent significant positive effects for all deciles, but as in the case of ELA, the decile estimates fall within the confidence interval of the OLS estimate for the full sample. In sum, Figures 2a, 2b and 2d show little difference in efficacy across student growth deciles while Figure 2c reveals a more pronounced benefit of *S-Portal Use* in the lower deciles of student math growth.

Recall that SGP is evaluated several months after the end-of-year CST tests are given. Thus, the above evidence does *not* suggest that teachers observe SGP and use PMSs in response. Rather, the teacher uses PMSs throughout the school year to detect progress against key learning objectives and devises intervention strategies to increase the likelihood of good CST and SGP results. The evidence suggests that, conditional on a student being in one of the lowest growth deciles for math, if the teacher makes greater use of the portal throughout the year, the student subsequently experiences significantly greater growth than a student in the *same* growth decile whose teacher makes less use of the portal. In short, the portal PMS is particularly helpful in fulfilling the organization's mission of targeting at-risk students.

5.3 Robustness tests with Teacher as the Unit of Analysis

Performance is measured at the student-level and the use of PMSs is measured at the teacher-level; however, students are not fully nested within teachers.[24] Earlier sections present associations between student outcomes and average PMS use of the student's teacher(s). This approach aggregates at most four teachers and provides a powerful test of the association between teacher actions and student outcomes. As a robustness test, this section presents teacher-level analysis of the association between individual teachers' use of the PMSs and aggregate performance of all students taught by the teacher.

For the three years of study, there are 625 teacher-years that are associated with 383 unique teachers, 309 of whom teach ELA and 314 of whom teach math. This translates to 505 teacher-years of math and 509 teacher-years of ELA. Teachers are associated on average with 28 (math)/ 27 (ELA) students (median of 22 for both subjects). For the teacher-level analysis, *ELAsgp* and *Mathsgp* are measured as the median performance of all students associated with a teacher. The two student demographic variables are measured as the percentage of a teacher's students in the category. Students may have up to four teachers each school year, so a specific student may appear in the performance outcomes for up to four teachers. As in the student analysis, we are also interested in how PMS use affects students in different parts of the SGP distribution; thus, we examine the association between PMS use and two additional moments of the SGP distribution: the 25th and the 75th percentiles.

Appendix Table A-1 provides descriptive statistics for the variables in the teacher-level analysis. The row labeled 'Between schools,' provides descriptive statistics for the 30 school-level means and indicates variation across schools. The row labeled, 'Within schools,' provides descriptive statistics for: (teacher-year value) – (school-level mean) + (overall mean), which indicates variation within schools. The data indicate significant variation in SGP and in PMS use both within and between schools; however, with one exception (*PctEngonly*), there is greater variation within than between schools. Naturally, aggregation of students creates a more compressed distribution as compared to the student-level data

[24] Although teachers and students are fully nested within schools, students share up to four teachers and even a set of teachers cannot be assumed to teach the same set of students. Consequently it is not possible to estimate a multi-level model in which students are nested within teachers, within schools.

(e.g., in Table 1 the standard deviation of *Mathsgp* for students is 27.0 while that of *Median Mathsgp* for teachers is 15.3). Nonetheless, there is still considerable variation in the moments of the distribution.

In untabulated analysis of univariate correlations, for the 240 teachers who teach both subjects, aggregate math and ELA SGP are correlated positively (approximately 0.46 for each moment of the distribution of students' growth). The two PMS use measures are positively correlated at a level of 0.29 ($p < 0.01$) and both are positively correlated with all moments of both subjects' SGP distributions. *Teacher effectiveness* is correlated with median *ELAsgp* and *Mathsgp* (0.30, 0.16, $p < 0.01$ for both) and the relation is similar for other moments of the SGP distributions. *Teacher effectiveness* is somewhat correlated (0.11 $p < 0.05$, .02 insignificant) with portal and Edusoft use, indicating possible moderate selection effects. Disadvantaged status (*PctDisadv*) is negatively associated with all moments of *ELAsgp* but has no association with *Mathsgp*. English-only status (*PctEngonly*) is positively associated with only the lowest *ELAsgp* moment. Absence of strong effects between student demographics and SGP that were evident in the student analysis indicates that students are relatively evenly distributed across teachers within a school. This is also evident in low 'Within School' variation in Table A-1.

Table 4, columns 1 and 2 report the results of OLS regression estimates for the relation:

$$\text{Median [math or ELA] SGP} = \beta_0 + \beta_1 (\textit{Portal Use}) + \beta_2 (\textit{Edusoft Use}) + \sum_{i=1}^{n} \beta i\ (Control\ i) \qquad (2)$$

where the control variables are the percentage of the teacher's students that are economically disadvantaged (*PctDisadv*) and that speak English only (*PctEngonly*), and the teacher's *effectiveness rating*. (Appendix Table A-2 reports results with the larger sample that is not reduced by availability of a teacher effectiveness rating). The models include untabulated significant model constants as well as jointly significant school fixed effects.[25]

[25] As a robustness check we estimate a multilevel model specification with teachers (Level-1) nested within schools (Level-2). The school effect is a random rather than a fixed effect. This is equivalent to the one-way ANOVA model (also termed the fully unconditional model). The base model with only school random effects significantly explains median student performance variation. Next we introduce level-1 predictors, *Portal Use* and *Edusoft Use*, and controls, while maintaining the school-level random intercept and constraining the effects of PMS use and the control variables to be fixed for all schools. The estimated coefficients are qualitatively similar to the OLS regression results. After including the independent variables, school accounts for 10 percent of the variation in *Median ELAsgp* and 7 percent of the variation in *Median Mathsgp*.

The results indicate that, as in the student analysis, *Median ELAsgp* is influenced positively by *Edusoft Use* ($p \leq 0.01$) but not *Portal Use,* controlling for teacher effectiveness and student demographics. Both *Edusoft-use* and *Portal Use* are associated with higher *Median Mathsgp*. Evaluated at the mean levels of *Edusoft Use* and *Portal Use* (Table A-1), *Median Mathsgp* is 5.9 points higher and *Median ELAsgp* is 3.4 points higher for teachers with average PMS usage. The results are broadly consistent with student-level results, with improved model fit.

As in the student analysis, we compute effect sizes on median SGP that are associated with PMS use and teacher effectiveness. For PMS use we consider three "treatment effects" defined by above-median use of the portal and Edusoft (BOTH), above-median use of the portal coupled with below-median use of Edusoft (PORTAL), and below-median use of the portal coupled with above-median use of Edusoft (EDUSOFT), all as compared with below-median use of both PMSs. For *mathsgp*, all three treatments are significantly different from zero (95 percent confidence interval), and the effect sizes (i.e. Cohen's *d*) are .62, .28, and .27, for BOTH, EDUSOFT and PORTAL. For *ELAsgp*, the BOTH treatment and the EDUSOFT treatment are significantly different from zero, and Cohen's *d* is 0.54 and 0.34, respectively. (Cohen's *d* for PORTAL is 0.18 and not significantly different from zero). The effect size on students' median SGP of changing teachers to move from teachers with a below- to an above-median teacher effectiveness rating is $d = 0.43$ for ELA and $d = 0.21$ for Math (both significant).[26]

In his review of studies of education innovations related to student achievement, Hattie (2009, 17) argues that an effect size of 0.40 is a threshold at which "…the effects of innovation enhance achievement in such a way that we can notice real-world differences." Between seven and 21 percent of the variance in student achievement is associated with teacher effectiveness, which corresponds to an effect size of $d = 0.32$ (Hattie 2009, 108). Our estimates of effect sizes for teacher effectiveness are similar, providing high external validity for the results. Since the PMS effect sizes compare above- to below-median use rather

[26] The larger effect sizes of the teacher-level analysis as compared with the student-level analysis reflects the greater variation in the student level data as compared with the teacher-level model that estimates the average performance of all students taught by the teacher. Simply put, it is easier to estimate average student performance than individual student performance.

than ‘use’ to ‘no use’, full effects of individual PMS use may well exceed Hattie’s threshold for effective innovations and taken together the effects of both PMSs exceed the threshold.

Continuing the inquiry into whether students with different growth realizations vary in responsiveness to teacher use of PMSs, the results of estimating equation (2) for the 25th and 75th percentile of aggregate student SGP are in the columns to the right of the median results in Table 4 (Table A-2 reports models that exclude *teacher effectiveness*). For *ELAsgp*, the results are broadly similar to the median results. *Edusoft use* is associated with similar levels of increased growth at all moments of the distribution. For *Mathsgp*, *Edusoft use* is uniformly associated with significant growth. *Portal Use* is positively associated with growth in both percentiles but, as in the student quantile results, has a greater impact in the 25th percentile.

5.4 Robustness tests: Endogeneity of PMS use and selection effects

In the previous analyses, *teacher effectiveness* is a control variable used to distinguish the effect of PMS use from the potentially correlated selection effect of more effective teachers using the PMSs. This presumes that the principal’s rating captures all information about the teacher’s decision to use the PMS that is relevant to student learning growth. In columns 3 and 4 of Table 4 we employ a difference-in-difference approach in which the teacher serves as her own control. Since teacher effectiveness is measured in (at most) the last two years, for comparability with columns 1 and 2 we use only teachers who are present for these two years. The results indicate that use of both PMSs is positively associated with median *Mathsgp* and use of the Edusoft PMS is positively associated with median *ELAsgp.*

In a final effort to disentangle the effects of PMS use from a selection effect, we use a simulation analysis approach known as a “shuffle test” (La Fond and Neville 2010, Anagnostopoulos, Kumar and Mahdian 2008). For each subject, we compare the actual correlation between the teacher’s median student SGP and use of the PMSs with a distribution of correlations under the null hypothesis that the PMSs have no effect beyond that associated with the selection bias. The test indicates with high reliability that correlations between SGP and the PMSs’ use are indeed associated with PMS use rather than with better teachers using the PMSs. Because it is somewhat novel, we describe the test more fully below.

We start from the premise that SGP and PMS use are correlated due to the hypothesized effect of PMS use on SGP (feedback effect) and the effect of teachers who consistently deliver high student growth making more use of the PMSs (selection effect). The test does not assume a directionality, level, or functional form for the selection effect. We assume that teacher ability is constant over short durations (i.e., within the three-year period) and that a teacher's PMS use in a year can only influence learning outcomes of students who she teaches that year (i.e., no spillovers to other students or years).

Teacher j's average student SGP and PMS use in two consecutive years is defined by:

$SGP_{(j,\ t+1)} = SGP_{(j,t)} + \text{deltaSGP}$; and (3)

$PMS_{(j,t+1)} = PMS_{(j,t)} + \text{deltaPMS}$ (4)

where deltaSGP and deltaPMS are the difference in variable levels between year *t+1* and *t* for teacher *j*. To disentangle the selection effect from the PMS effect, we estimate the distribution of the correlation between PMS and SGP that is expected under the null hypothesis that PMS use has no effect on SGP and that correlation is due only to the selection effect. To facilitate this, we construct a new measure, $SGP'_{(j,t+1)}$, in which we replace deltaSGP in equation (3) with the deltaSGP of another randomly chosen teacher of the same subject.[27] Randomizing the values for deltaSGP across teachers leaves intact the impact of the selection effect because selection information is recorded in the teacher's choice of how much to use the PMS (e.g., retaining the common influence of teacher ability on PMS use) and breaks the hypothesized causal link whereby a changed level of PMS use yields a changed level of SGP.

We generate a distribution for the correlation between the new SGP' (which includes the randomized element) and PMS by running 10,000 Monte Carlo repetitions of the deltaSGP randomization procedure for each subject, for each PMS. This is the distribution of the SGP' and PMS use correlation

[27] We use the 'simulate' command in Stata to randomly select without replacement a deltaSGP value from the set of all actual deltaSGP values of teachers who teach the subject (ELA or math). With each repetition, this command shuffles the deltaSGP values, produces a value for SGP', and records the correlation between SGP' and PMS use. We use these 10,000 values to estimate the distribution of the correlation under the null hypothesis. This test is only possible for the subset of teacher-year observations for which we have a prior year for comparison (N=200 for ELA, N=191 for math). The results are robust to changes in a teacher's students from one year to the next because the SGP measure always compares students against their individually tailored peer group.

under the null hypothesis that there is no feedback effect. It is a short leap to see that a true correlation between SGP and PMS that falls in the upper tail of the distribution of the correlation between SGP' and PMS indicates a feedback effect that is separate from any selection effect. Indeed this is what is found. The simulations produce a distribution of the correlation between SGP' and *Edusoft use* with mean 0.087, median 0.087, and standard deviation 0.051 for ELA, and mean 0.114, median 0.114, and standard deviation 0.054 for math. The simulated distribution of correlation between SGP' and *Portal use* has mean 0.108, median 0.107, and standard deviation 0.051 for ELA, and mean 0.127, median 0.127, and standard deviation 0.053 for math. Returning to the (non-randomized) data, the actual correlation between median SGP and *Edusoft use* is 0.188 and 0.170 for ELA and math, respectively, and the actual correlation between median SGP and *Portal use* is 0.063 and 0.206 for ELA and math, respectively. With the exception of the effect of *Portal use* on ELAsgp, these figures are significant at the 10% level (one-tailed test), when compared with the simulated distribution. Thus, with high probability we can conclude that the documented association between PMS use and SGP is attributable to PMS use rather than an artifact of high-performing teachers making systematically different choices about PMS use.

6. Conclusion

This study contributes important evidence to the management accounting literature on the decision-facilitating role of PMSs. Experimental studies show that PMSs influence decision-makers' choices. Field and survey research provide evidence linking the presence and self-reported use of PMSs to organizational performance, but are largely silent on whether reported associations are driven by selection effects rather than the hypothesized feedback effect. This study bridges the different empirical approaches to link real decision-makers' actual *use* of PMSs to organizational performance, finding that the predicted positive association is statistically and materially significant, and distinct from selection effects associated with decision to use the PMS. We demonstrate the potential for using unobtrusive data on PMS use to probe the mechanisms of performance improvement in organizations and to link the collective actions of individual decision-makers to organizational outcomes.

The results suggest a direction for future research into pervasive ‘manager fixed effects’. In the education sector it is well-established that teacher quality is the single most influential factor in student performance that schools control (e.g., Hanushek and Rivkin 2010; Hanushek 2011). However, “only about three percent of the contribution teachers made to student learning was associated with teacher experience, degree attained, and other readily observable characteristics. The remaining 97 percent of their contribution was associated with *qualities or behaviors that could not be isolated and identified* (Goldhaber 2002, 3, italics added).” This study examines one classroom behavior, teachers’ use of feedback from PMSs, and finds that teachers who use PMSs to identify student learning difficulties and plan targeted interventions experience greater student performance improvement. Just as teachers’ use of PMS feedback to diagnose and intervene in student performance improves student growth, manager fixed effects that are defined by delivering better financial outcomes (Demerjian, Lev and McVay, 2012) may be associated with more effective use of PMS feedback to improve organizational performance.

ASPIRE is a “data-driven organization,” as its teachers and principals openly acknowledge and as school bulletin boards and the organization’s web site proclaim. We have investigated selection effects of PMS use within ASPIRE; however, the organizational culture undoubtedly creates selection bias in employing those who share a conviction about the value of data analysis. If ASPIRE attracts teachers who are more productive in the use of PMS and gain greater marginal benefits per unit of “PMS use” than teachers in another organization, then the results may not hold in another school system. We found no evidence that ASPIRE uses data analysis skill as a hiring criterion; however, self-selection of employees into preferred work environments could create this effect. In counterpoint, the teachers exhibit very significant variation in their use of the PMSs; thus, any selection bias related to employment and retention does not yield uniformly high PMS use.

Finally, there are caveats related to the translation of theory into empirical tests. The theory behind feedback and formative assessment is that decision-makers observe intermediate (“leading indicator”) data, diagnose difficulties, develop and administer remediation strategies, and that these steps deliver superior performance. Experimental studies have demonstrated that PMSs influence the diagnosis/remediation choices of decision-makers; however, these studies cannot link decision-makers' choices to organizational outcomes. This paper focuses on the connection between actual use of PMSs and organizational performance. Unlike a controlled experiment in which the content and presentation of the intermediate information may be manipulated to assess what diagnostic responses are induced, in the natural field setting the range and complexity of intermediate information is extensive and unspecified. Thus, although we measure actual exposure to PMS data, we have only anecdotal interview evidence of how teachers use the PMS data to improve student performance. The study provides a unique window on the 'decision-facilitating' role of PMSs as compared to prior research; however, much remains to be learned about how decision-makers use PMS data to generate superior performance.

REFERENCES

Anagnostopoulos, A., Kumar, R., and Mahdian, M. (2008, August). Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 7-15). ACM.

Bennett, R. E. 2011. Formative assessment: A critical review. *Assessment in Education: Principles, Policy and Practice*. 18 (1): 5-25.

Betebenner, D. W. 2008. A Primer on Student Growth Percentiles. Working paper. National Center for the Improvement of Educational Attainment. Feb. 8, 2008.

Black, P. and D. Wiliam. 1998. Assessment and Classroom Learning. *Assessment in Education*. 5 (1): 7-74.

Cardinaels E., and P. Van Veen-Dirks. 2010. Financial versus non-financial information: the impact of information organization and presentation in a Balanced Scorecard. *Accounting, Organizations and Society*, 35 (6): 565-578.

Casas-Arce, P. F.A. Martinez-Jerez, and V.G. Narayanan. (2017) The impact of forward-looking metrics on Employee decision-making: The case of Customer Lifetime Value. *The Accounting Review*. 92 (3): 31-56.

Cheng, M. M., and K. A. Humphreys. 2012. The differential improvement effects of the strategy map and scorecard perspectives on managers' strategy judgments. *The Accounting Review*. 87 (3): 899-924.

Chetty, R., J. N. Friedman, and J. E. Rockoff. 2014a. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review* 104(9): 2593-2632

Chetty, R., J. N. Friedman, and J. E. Rockoff. 2014b. Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review* 104(9): 2633-2679.

Childress, S., R. Elmore, A. Grossman, and S. Johnson, Eds. 2007. *Managing School Districts for High Performance*. (Harvard Education Press: Cambridge MA)

Davis, S., and T. Albright. 2004. An investigation of the effect of the balanced scorecard implementation on financial performance. *Management Accounting Research*. 15 (2): 135-153.

Demerjian, P., B. Lev, and S. McVay. 2012. Quantifying managerial ability: A new measure and validity tests. *Management Science* 58(7), 1229-1248.

Demski, J., and G. A. Feltham. 1976. *Cost Determination: A conceptual approach*. (Ames, IA: Iowa State University Press).

DiMaggio and Powell 1983. The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*. 48 (2): 147-160.

Dunn, K.E., D. Airola, W-J Lo, and M. Garrison. 2013. Becoming data driven: The influence of teachers' sense of efficacy on concerns related to data-driven decision making. *Journal of Experimental Education,* 81 (2): 222-241.

Ehlert, M., C. Koedel, E. Parsons, and M. Podgursky. 2012. Selecting growth measures for school and teacher evaluations. National Center for Analysis of Longitudinal Data in Education Research. Working paper 80. August.

Farrell, A. M., K. Kadous, and K. L. Towry. 2012. Does the communication of causal linkages improve employee effort allocations and firm performance? An experimental investigation. *Journal of Management Accounting Research*. 24: 77-102.

Fensterwald, J. 2015a. “Panel recommends continuing districts’ waiver from NCLB” EdSource. June 25. http://edsource.org/2015/panel-recommends-continuing-districts-waiver-from-nclb/82052

Fensterwald, J. 2015b. “No child left behind waiver extended for CORE districts” EdSource. September 25. http://edsource.org/2015/no-child-left-behind-waiver-extended-for-core-districts/87664.

Firestone, W.A. and J.R. Pennell. 1993. Teacher Commitment, Working Conditions, and Differential Incentive Policies. *Review of Educational Research*. 63 (4): 489-525.

Fryer, R.G. 2013. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*. 31 (2) pt.1.: 373-406.

Goldhaber, D. 2002. The Mystery of Good Teaching. *Education Next*, 2 (1): 50-55.

Grafton, J., A. Lillis and S. Widener. 2010. The role of performance measurement and evaluation in building organizational capabilities and performance. *Accounting Organizations and Society*. 35: 689-706.

Grissom, R. and J. Kim. 2012. *Effect sizes for research: univariate and multivariate applications*. Taylor & Francis Group LLC. New York, NY.

Hanushek, E. 2006. Teacher Quality. In Hanushek and Welch (Eds.) *Handbook of the Economics of Education*, 2, 1052-1078.

Hanushek, E. 2011. The economic value of higher teacher quality. *Economics of Education Review*. 30: 466-479.

Hanushek, E., and S. Rivkin. 2010. Generalizations about using value-added measures of teacher quality. *American Economic Review: Papers and Proceedings*. 100 (May): 267-271.

Hattie, J. 2009. *Visible Learning*. NY, NY: Routledge.

Henry, G., C. L. Thompson, C. K. Fortner, R. A. Zulli, and D. C. Kershaw. 2010. The Impact of Teacher Preparation on Student Learning in North Carolina Public Schools. White paper UNC at Chapel Hill.

Humphreys, K., and K. Trotman. 2011. The balanced scorecard: The effect of strategy information on performance evaluation judgments. *Journal of Management Accounting Research*. 23 (1): 81-98.

Ikemoto, G. and J. Marsh 2007. Cutting through the “data-driven” mantra: Different conceptions of dat-driven decision making. In P.A. Moss (Ed.), *Evidence and decision making*. Malden, MA: Blackwell Publishing Inc. Vol. 106: 105-131.

Ittner, C., D. Larcker and T. Randall. 2003. Performance implications of strategy performance measurement in financial services firms. *Accounting Organizations and Society*. 28: 715-741.

Kaplan, R. S., and D. P. Norton. 2004. *Strategy maps: Converting intangible assets into tangible outcomes*. Boston: Harvard Business School Press.

Kaplan R. S., and M. N. Lee. 2007 Fulton County School System: Implementing the Balanced Scorecard. *Harvard Business School Case* 107-029. January (revised August 2007).

Koenker, R., and K. F. Hallock. 2001. Quantile Regression. *The Journal of Economic Perspectives*. 15 (4): 143-156.

Keuning, T., M. Van Geel, and A. Visscher. 2017. Why a data-based decision-making intervention works in some schools and not in others. *Learning Disabilities Research and Practice*. 32 (1). 32-45.

La Fond, T. and J. Neville. 2010, April. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th international conference on World Wide Web*, (pp. 601–610). ACM.

Libby, T., S. Salterio, and A. Webb. 2004. The balanced scorecard: The effects of the assurance and process accountability on managerial judgment. *The Accounting Review*. 79 (4): 1075-1094.

Lipe, M., and S. Salterio. 2000. The balanced scorecard: Judgmental effects of common and unique performance. *The Accounting Review*. 75 (3): 283-298.

Malina, M., and F. Selto. 2001. Communicating and controlling strategy: An empirical study of the effectiveness of the balanced scorecard. *Journal of Management Accounting Research*, 13: 47-90.

McKinsey & Co. 2010. How the world's most improved school systems keep getting better. White paper.

Meyer and Rowan 1977. Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*. 83 (2): 340-363.

Rethinam, V. 2014. Predictive Analytics in K-12: Advantages, Limitations and Implementation. *T.H.E. Journal*. June 12.

Rich, M. 2013. Reading Gains Lag Improvements in Math: In Raising Scores, 1 2 3 is Easier than A B C. *New York Times*. May 29, 2013.

Rigby, D., and B. Bilodeau. 2013. *Management Tools and Trends 2013*. Bain & Company white paper. www.bain.com/publications/articles/management-tools-and-trends-2013.aspx

Rivkin, S., E. Hanushek and J. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica*. 73 (2) 417-458.

Slavin, R., A. Cheung, G. Holmes, N. Madden and A. Chamberlain. (2013) Effects of a data-driven district reform model on state assessment outcomes. *American Educational Research Journal*. 50 (2): 371-396.

Sprinkle, G. 2003. Perspectives on experimental research in managerial accounting. Accounting, Organizations and Society. 28: 287-318.

Staman, L., A.C. Timmermans and A.J. Visscher. 2017. Effects of a data-based decision making intervention on student achievement. *Studies in Educational Evaluation*. 55: 58-67.

Stecker, P.M., L.S. Fuchs and D. Fuchs. 2005. Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, 42 (8): 795-819.

Stiggins, R. 2005. From formative assessment to assessment for learning: A path to success in standards-based schools. *Phi Delta Kappan*, 87 (4): 324-28.

Stronge, J., T. Ward and L. Grant. 2011. What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*. 62 (4) 339-355.

Taylor, W. B. 2010. The balanced scorecard as a strategy-evaluation tool: The effects of implementation involvement and a causal-chain focus. *The Accounting Review*. 85 (3): 1095-1117.

Tuttle, C. C., B. Gill, P. Gleason, V. Knechtel, I. Nichols-Barrer and A. Resch. 2013. *KIPP Middle Schools: Impacts on Achievement and Other Outcomes*. Mathematica Policy Report. Ref number 06441.910. February 27, 2013.

Vera-Munoz, S., M. Shackell and M. Buehner. 2007. Accountants' usage of causal business models in the presence of benchmark data: A note. *Contemporary Accounting Research*. 24 (3): 1015-1038.

Visscher, A. and R. Coe. 2003. School performance feedback systems: conceptualization, analysis, and reflection. *School effectiveness and school improvement*. 14 (3): 321-349.

Table 1: Descriptive Statistics

Panel A: Variable descriptive statistics

VARIABLE NAME (Number of unique Observations)	**MEAN**	**STD DEV**	**MIN**	**MAX**
Dependent Variable: *ELAsgp* (12,649)	50.5	28.6	1.0	99.0
Independent Variables:				
S-Portal Use by ELA teacher(s) (12,649)	31.1	33.3	0.0	269.0
S-Edusoft Use by ELA teacher(s) (12,649)	161.3	118.5	0.0	753.0
Dependent Variable: Mathsgp (12,618)	50.5	27.0	1.0	99.0
Independent Variables:				
S-Portal Use by Math teacher(s) (12,618)	30.5	33.9	0.0	269.0
S-Edusoft Use by Math teacher(s) (12,618)	169.6	121.0	0.0	753.0
Control Variables				
Avg Teacher Effectiveness (292)	87.4	15.6	21.0	100.0
English only (6515)	0.48	.50	0.00	1.00
Disadvantaged (6515)	0.80	.40	0.00	1.00

Variable definitions:

Mathsgp, ELAsgp: the student's SGP performance for the year for math and ELA, respectively

S-portal Use, S-edusoft Use: the average use of each data tool by all of a student's teachers of the specified subject (up to four) in a school year

Avg Teacher Effectiveness: Each teacher's average principal rating for 2012-13 and 2013-14 is computed using available data. For the student-level analysis, the effectiveness of all teachers (up to four) who teach a student in a particular year are averaged. There are 292 unique combinations of teachers who teach students during the three years. Missing data on teacher ratings reduces the sample of students to 10,208 student-years of observations for ELA and 9,907 for Math.

English only: indicator variable set to 1.0 if the student is identified as speaking only English

Disadvantaged: indicator variable set to 1.0 if the student is from an economically disadvantaged home

Panel B: Correlation Tables with pairwise deletion of cases

	ELAsgp N=12,649	**S-Portal use** N=12,649	**S-Edusoft use** N=12,649	**Avg Teacher Effectiveness** N=10,208	**English only** N=12,649
S-Portal Use	0.02 *				
S-Edusoft Use	0.03 **	0.18 ***			
Avg Teacher Effectiveness	0.07 ***	0.12 ***	-0.04 ***		
English only	0.01	0.01	-0.13 ***	0.08 ***	
Disadvantaged	-0.05***	0.03 ***	0.16 ***	-0.06 ***	-0.36 ***

	Mathsgp N=12,618	**S-Portal use** N=12,618	**S-Edusoft use** N=12,618	**Avg Teacher Effectiveness** N=9,907	**English only** N=12,618
S-Portal Use	0.09 ***				
S-Edusoft Use	0.09 ***	0.18 ***			
Avg Teacher Effectiveness	0.10 ***	0.12 ***	-0.04 ***		
English only	0.01	0.01	-0.13 ***	0.08 ***	
Disadvantaged	-0.04 ***	0.03 ***	0.16 ***	-0.06 ***	-0.36 ***

***, **, * *p*-value ≤ 0.001, 0.01, 0.05, two tailed test of significance

Table 2: OLS Regression Analysis

The table presents OLS regression analysis of the association between student growth (SGP) and teachers' use of the two PMSs. Tests of significance employ robust standard errors clustered on students with a common teacher or teacher set. The first model for each subject includes only controls for student demographics. The second model for each subject includes the *average teacher effectiveness* control variable. This model has a reduced sample size due to missing data on teacher ratings, but sheds light on the separate effects on SGP of teacher effectiveness and teachers' use of PMSs. Untabulated fixed school effects (30 schools) are significant in both models. Significant model constants are not reported to preserve confidentiality of the data.

Variable Name	***Predicted Sign***	**ELAsgp** ***Estimated coefficient*** ***(t-statistic)***		**Mathsgp** ***Estimated coefficient*** ***(t-statistic)***	
		N=12,649 Std Errors adjusted for 311 teacher combinations	N=10,208 Std Errors adjusted for 232 teacher combinations	N=12,618 Std Errors adjusted for 332 teacher combinations	N=9,907 Std Errors adjusted for 247 teacher combinations
S-Portal Use	(+)	0.019 (1.45) *	-0.012 (-0.98)	0.055 (3.01) ***	0.051 (2.46) ***
S-Edusoft Use	(+)	0.014 (2.68) ***	0.013 (2.75) ***	0.022 (2.78) ***	0.017 (1.81) **
English only		-1.06 (-1.76) *	-0.987 (-1.49)	-1.59 (-2.64) ***	-2.20 (-3.20) ***
Disadvantaged		-2.72 (-4.21) ***	-2.04 (-2.93) ***	-2.67 (-3.16) ***	-2.87 (-3.06) ***
Avg Teacher Effectiveness			0.231 (5.96) ***		0.208 (2.75) ***
Adjusted R^2		0.030	0.042	0.046	0.055

***, **, * p-value ≤0.01, 0.05, 0.10 in one-tailed test of significance for independent variables and two-tailed test for controls.

Table 3: Quantile Regression Analysis

Quantile regression of the association between student growth and use of the diagnostic tools for deciles of the SGP distribution. For the sake of parsimony, we tabulate only the coefficients for the two PMSs, *S-Portal Use* and *S-Edusoft use*; however, the estimated model includes the same variables as the models of Table 2 that include *teacher effectiveness*. We use bootstrapped standard errors to obtain confidence intervals for Figure 2.

	ELAsgp **N=10,208**				**Mathsgp** **N=9,907**			
Independent Variable:	**S-Portal Use**		**S-Edusoft Use**		**S-Portal Use**		**S-Edusoft Use**	
	Coeff.	**t-stat**	**Coeff.**	**t-stat**	**Coeff.**	**t-stat**	**Coeff.**	**t-stat**
OLS regression (repeated from Table 2)	-0.012	-0.98	0.013	2.75 ***	0.051	2.46 ***	0.017	1.81 **
Decile:		**z-stat**		**z-stat**		**z-stat**		**z-stat**
0.10	0.005	0.35	0.008	1.68 **	0.051	2.28 **	0.013	1.93 **
0.20	0.006	0.30	0.019	2.72 ***	0.071	2.70 ***	0.020	2.08 **
0.30	-0.017	-0.94	0.019	2.19 **	0.082	2.73 ***	0.022	2.00 **
0.40	-0.025	-1.14	0.020	2.17 **	0.078	2.37 ***	0.023	1.82 **
0.50	-0.032	-1.62	0.016	1.93 **	0.068	2.13 **	0.017	1.32 *
0.60	-0.020	-0.93	0.013	1.69 **	0.057	1.87 **	0.015	1.19
0.70	-0.029	-1.76	0.016	2.40 ***	0.044	1.62 *	0.016	1.41 *
0.80	-0.020	-1.44	0.010	1.76 **	0.024	1.07	0.014	1.37 *
0.90	-0.004	-0.39	0.007	1.75 **	0.011	0.64	0.012	1.45 *
Avg. Pseudo R^2	0.026				0.035			

***, **, * p-value ≤0.01, 0.05, 0.10 in one-tailed test of significance.

Table 4: Teacher-level Analysis

Columns [1] and [2] report results of OLS regression analysis of the association between median, 25th percentile and 75th percentile SGP performance for all students taught by a teacher, and the teacher's use of PMSs, controlling for student demographics and teacher effectiveness. Untabulated fixed school effects (N=30 schools) and a model constant are significant in all models. (a) Columns [3] and [4] estimate a difference-within-difference model in which year to year change in median SGP is regressed against change in PMS use, controlling for changes in demographics of the students taught in the two years. In this model, the teacher acts as her own control, so the teacher effectiveness measure is removed and because teachers do not change schools, school fixed effects are removed.

	Predicted Sign	**Column [1]** **ELAsgp** N=375 ***Estimated coefficient*** ***(t-statistic)***			**Column [2]** **Mathsgp** N=365 ***Estimated coefficient*** ***(t-statistic)***			**Col [3]** **Delta** **Median** **ELAsgp** **N=92**	**Col [4]** **Delta** **Median** **Mathsgp** **N=90**
Independent Variable		**Median**	**25th**	**75th**	**Median**	**25th**	**75th**		
Portal Use	(+)	-0.025 (-1.37)	-0.018 (-1.04)	-0.002 (0.10)	0.049 (2.27) **	0.062 (2.75) ***	0.039 (2.10) **	-0.001 (-0.02)	0.081 (1.85) **
Edusoft Use	(+)	0.017 (2.30) **	0.019 (2.77) ***	0.012 (1.91) **	0.021 (2.74) ***	0.019 (2.45) ***	0.017 (2.29) **	0.019 (1.34) *	0.036 (2.88) ***
Controls									
PctEngonly		-12.9 (-1.70) *	-11.2 (-1.38)	-7.4 (-1.18)	-20.8 (-2.38) **	-15.2 (-1.74) *	-22.3 (-2.77) ***	-24.4 (-2.00) **	-10.1 (-0.61)
PctDisadv		-10.6 (-1.17)	-15.3 (-1.81) *	-18.7 (-2.74) ***	4.25 (0.47)	0.94 (0.09)	12.3 (1.46)	-8.3 (-0.43)	-34.6 (-1.31)
Teacher Effectiveness		0.278 (5.71) ***	0.275 (6.70) ***	0.282 (6.32) ***	0.207 (4.01) ***	0.229 (4.65) ***	0.161 (3.42) ***		
Adjusted R^2		0.20	0.19	0.25	0.13	0.13	0.11	0.02	0.08

***, **, * p-value ≤0.01, 0.05, 0.10 in one-tailed test of significance for independent variables and two-tailed test for controls.

(a) In untabulated results of maximum likelihood estimation of a two-level model with random school effects, school-level variation (i.e., rho) explains 14, 10, 14 percent of 25th percentile, median, 75th percentile *ELAsgp* variation and 5, 7, 5 percent of 25th percentile, median, 75th percentile *Mathsgp* variation.

Figure 1: Frequency Distribution of students’ California Standardized Test (CST) Scores for ELA and Math by year

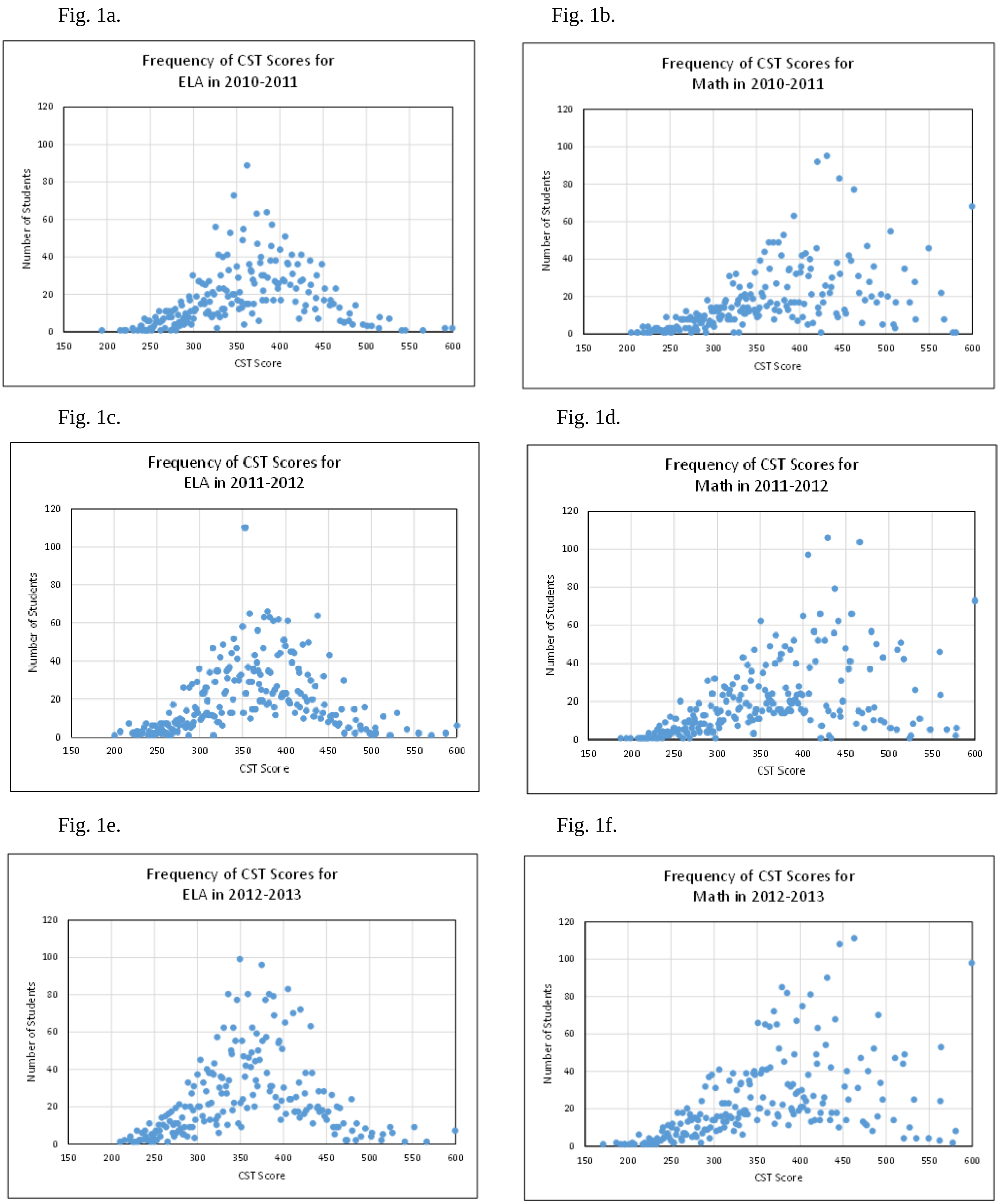

Fig. 1a.

Fig. 1b.

Fig. 1c.

Fig. 1d.

Fig. 1e.

Fig. 1f.

Figure 2: Comparison of OLS coefficient and quantile regression estimates of the marginal effects on SGP of the use of PMSs for deciles of the SGP distribution.

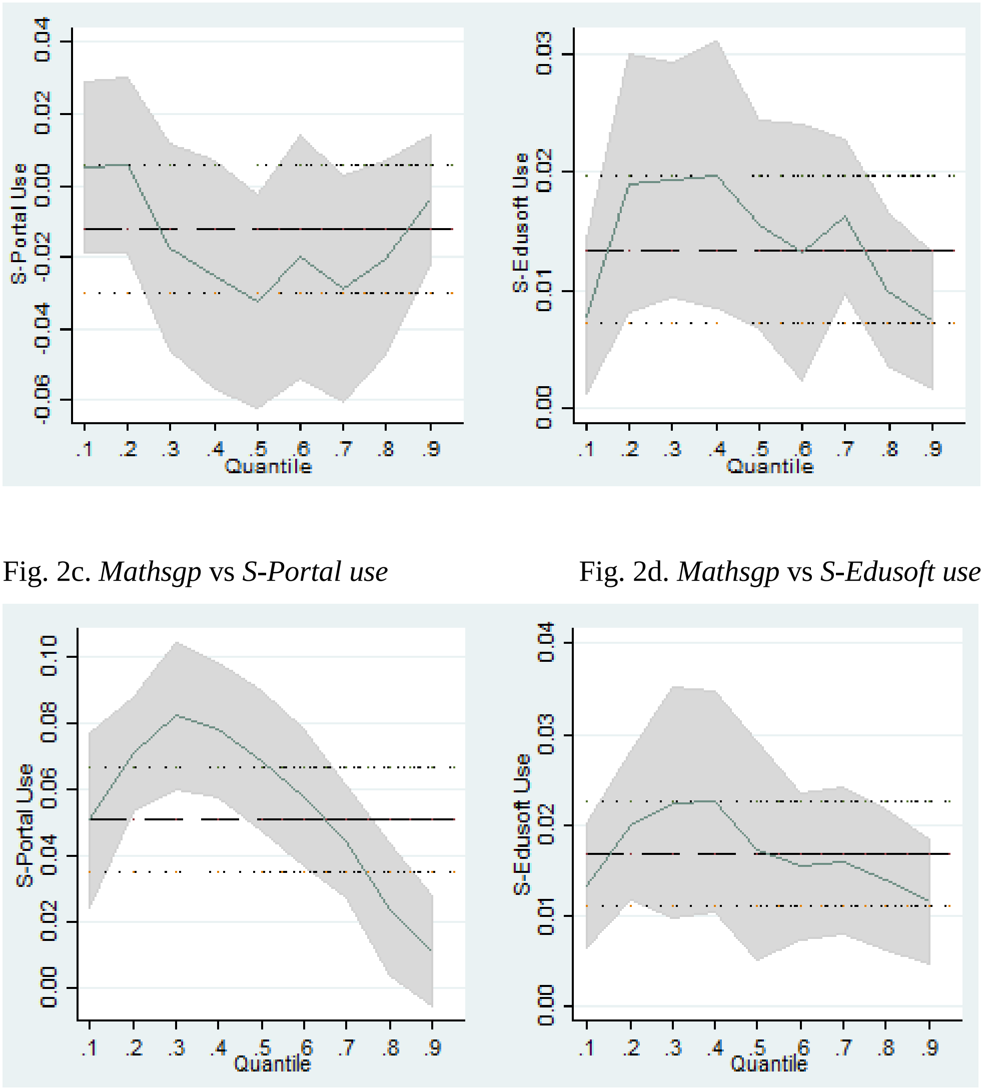

Fig. 2a. *ELAsgp* vs *S-Portal use*

Fig. 2b. *ELAsgp* vs *S-Edusoft use*

Fig. 2c. *Mathsgp* vs *S-Portal use*

Fig. 2d. *Mathsgp* vs *S-Edusoft use*

Notes: For each variable, the horizontal dashed line represents the OLS estimated coefficient and the two dotted lines on either side are the 90 confidence intervals for the least squares estimate. The solid line connects the point estimates of the coefficient from quantile regressions and the shaded area around it is a 90 percent confidence band for these estimates.