Making Time for Instructional Leadership

VOLUME 2: THE FEASIBILITY OF A RANDOMIZED CONTROL TRIAL OF THE SAM PROCESS

Ellen Goldring, Jason A. Grissom, Christine M. Neumerski
Joseph Murphy, Richard Blissett VANDERBILT UNIVERSITY

Andy Porter UNIVERSITY OF PENNSYLVANIA

VANDERBILT PEABODY COLLEGE

The Wallace Foundation®

This study was commissioned by The Wallace Foundation. It was conducted by researchers at Vanderbilt University and the University of Pennsylvania. The contents expressed in this report are those of the authors and do not necessarily represent the views of the sponsor.

Cover photo: vgajic
Front and back cover design: José Moreno

Acknowledgement

The support and valuable contributions of several individuals and organizations were essential to this study. We thank Mark Shellinger, Director, National SAM Innovation Project, for his ongoing engagement and valuable feedback throughout this project. Jim Mercer at NSIP assisted us with survey implementation and was consistently helpful with our numerous data questions. Other staff from NSIP gave generously of their time and provided crucial insights. We thank the participants in this study who welcomed the researchers into their schools and districts, and shared their experiences and expertise through interviews and surveys. Mollie Rubin and Laura Rogers provided valuable research support.

MAKING TIME FOR INSTRUCTIONAL LEADERSHIP
VOLUME 2: THE FEASIBILITY OF A RANDOMIZED CONTROLLED TRIAL OF THE SAM PROCESS

Ellen Goldring
Jason A. Grissom
Christine M. Neumerski
Joseph Murphy
Richard Blissett
Vanderbilt University

Andy Porter
University of Pennsylvania

Table of Contents

I. Purpose of This Report and Approach

A goal of this project was to collect and assess information that would shed light on whether the SAM® process is ready for a randomized controlled trial (RCT), whether a high-quality RCT is feasible, and, if so, to make recommendations regarding what steps might be taken to facilitate a high-quality RCT. This appendix discusses findings on each of those fronts.

We begin by discussing the five criteria we identified for assessing the feasibility of an RCT and an accompanying implementation analysis. Then, drawing on the data we collected to inform our investigation of the SAM process as detailed in the main report, we discuss what we learned in light of each of these criteria. The final section details our recommendations for designing a high-quality RCT and implementation analysis of the SAM process.

To preview our conclusions, there is considerable support for conducting an RCT of the SAM process. The study data suggest that there is strong fidelity of implementation and that the SAM process can be implemented at sufficient scale for an RCT. Furthermore, our data to suggest efficacy of the SAM process in changing behaviors or outcomes consistent with its theory of action, including changes in instructional time use. There are sufficient existing measures to systematically study the proximate and ultimate outcomes of the SAM process, and it is feasible to design and implement a rigorous RCT design if attention is given to the crucial design parameters and considerations we describe.

II. Criteria for Judging the Feasibility and Utility of an RCT and Implementation Analysis

Our analysis and recommendations regarding the feasibility and utility of an RCT of the SAM process are based on five criteria that reflect those used by the U.S. Department of Education's Institute of Education Sciences in evaluating proposals for what they refer to as "Effectiveness" studies, which typically are large-scale RCTs of well-developed programs or interventions. The criteria are:

1. Are schools implementing the SAM process with fidelity?
2. Is the SAM process sufficiently well established and articulated that it can be replicated across a large number of schools in an RCT?
3. Does the SAM process show evidence of efficacy in changing behaviors or outcomes consistent with its theory of action?
4. Do measures exist (or could they reasonably be developed) to capture elements of the theory of action and hypothesized outcomes?
5. Is an RCT of the SAM process feasible from a design perspective? That is, is there a reasonable design for randomization and study implementation that could produce inferences regarding the impacts of the SAM process that are internally and externally valid and statistically precise?

The first criterion asks whether schools that are implementing the SAM process are doing so consistently as intended. Evidence that current participants are not engaging with the process

with fidelity would suggest that subjects in an RCT are unlikely to implement the process as it is designed.

Relatedly, the second criterion considers whether the SAM process can be taken to scale in an experiment in the sense that its components are stable and described well enough that subjects in an RCT could replicate the SAM process across many schools. If the SAM process is evolving rapidly, for example, schools may be implementing many variations on the process, making the program not replicable in the sense required by an RCT.

The third criterion asks whether existing evidence provides sufficient evidence of positive effects of the SAM process on outcomes it should affect to warrant a significant investment in further evaluation. These outcomes can be both *proximal* to the SAM process, such as by changing how principals allocate their time between instructional activities and other activities, and more *distal* outcomes, such as changes in the quality of teacher instruction as a result of changes in proximal outcomes. Minimal evidence of positive effects on relevant outcomes would provide little promise that positive effects would be uncovered in an RCT.

The fourth criterion explores whether the components of the SAM process and its hypothesized effects could be measured in a valid way. These effects include both medium-run or long-run *(final)* outcomes of participating in the SAM process and *mediating outcomes,* or near-term effects of the intervention that lead to later effects. For example, the SAM process might change principal time allocations, which then change teacher behaviors, such as instructional practices, which improve student achievement. If important components of the theory of action or the most relevant mediating or final outcomes could not be captured adequately without an unrealistically costly investment in data collection, an RCT would be difficult to justify.

The final criterion considers whether there is a logistically feasible design for an RCT that would permit convincing estimates of effects. There are numerous facets to consider. The first is *internal validity.* Can schools realistically be randomly assigned to the SAM process or to a business-as-usual control such that the control represents the desired counterfactual and the effects estimated are interpretable as due to the SAM process? Can the RCT be implemented in a way that guards against potential threats to validity that might be present even with random assignment, such as "spillover" among principals in the treatment and control conditions? The second is *precision* (or statistical power). Can an RCT be designed to be sufficiently precise to detect an effect size of educational importance on outcomes of interest? The third is *external validity,* or the extent to which the results of an RCT can be generalized. The technical way to explore external validity is to build a design that tests interactions, which in the case of the SAM process undoubtedly include variables such as principal experience, leader behaviors, and indicators of implementation, identified and addressed in the fieldwork component of our study.

We summarize in the following sections what we have learned about the SAM process in light of these criteria and conclude that the SAM process warrants the development of an RCT.

Criterion 1: Are Schools Implementing the SAM Process with Fidelity?

We conclude that the first criterion is met. As detailed in the main report, a large number of schools across numerous locales and types are implementing the SAM process with relative fidelity as evidenced by case study data, TimeTrack calendar data, and survey results,

notwithstanding local adaptation. Schools and districts are largely committed to the process, and core components of the process are evident in the schools. Principals are voluntarily participating with district buy-in.

From the survey data, we learned that the large majority of principals usually or always meet with their SAM, use and reconcile the TimeTrack calendar, hold reflective conversations, use the First Responder system, and meet with a Time Change Coach. The case studies further document the widespread implementation of the core components.

Throughout the report, we note variation and challenges in implementation as well. More nuanced and complex aspects of the process are most challenging and are implemented with less frequency and robustness. For example, the First Responder system presents challenges for some schools, and principals are much more likely to create their schedules and reconcile their calendars using the TimeTrack calendar than they are to use the calendar to disaggregate data to see how they implement specific tasks with specific teachers. Other challenges noted in the case studies refer to choosing an appropriate person to be a SAM and the extent to which the SAM is comfortable with and able to sufficiently probe and challenge the principal in the Daily Meetings. We note instances of lax use of the TimeTrack calendar, resulting in unscheduled time that we do not fully understand.

However, we conclude that the variation and challenges are within the realm of what one would expect for implementation of a complex behavioral intervention in schools that honors the need for local adaptation. Given the spread of the SAM process, across diverse locales, types of schools, districts, and principals, we conclude that there is a very high level of common implementation of the SAM process. There is a clearly defined, recognizable program that is sufficiently similar in its implementation to warrant an RCT.

This notion of fidelity with local adaptation is not unique, and it is discussed in the literature. Fidelity is commonly conceptualized as how closely implemented versions of programs align with original design (O'Donnell, 2008). Fidelity and adaptation are not mutually exclusive. Adaptation supporters argue that local conditions influence how practitioners implement programs; adaptation is necessary for successful implementation (Berman & McLaughlin, 1978).

Criterion 2: Is the SAM process sufficiently well established and articulated that it can be replicated across a large number of schools in an RCT?

We find that the second criterion is largely met. The SAM process is very clearly articulated and well understood. There is widespread agreement about the core components of the program and a well-defined theory of action. The participants in our study had no misunderstandings about the program components, expectations, or hypothesized benefits. Those engaged in the process, from NSIP personnel to district and school participants, each had a clear understanding of the core components and the respective roles and responsibilities of each party. NSIP has clear job descriptions, procedures for implementation and training, and ongoing support and professional development (PD).

The specificity of the SAM process and its theory of action bode well for a rigorous evaluation through an RCT. First, we believe that the program can be replicated and implemented with fidelity across a large number of schools because (a) there is a very clear articulation of the SAM

process and the core components and (b) there is a systematic approach to implementation and support already developed by NSIP. Virtually all of the principals who responded to our survey indicated they worked with an Implementation Specialist to learn about the SAM process. Respondents were similarly positive about the specific aspects of the training to learn the details of the SAM process. Furthermore, the Time Change Coaches, national conference, and PD workshops provide a support for implementation and consistency across schools. Ninety percent of principal respondents noted that they work with a Time Change Coach, and 74% found them to be "very helpful." Through our case studies, we learned that coaching support was also very valuable to SAMs and helped them develop, as well.

Second, the theory of action provides the necessary conceptualization of the mechanism, or the ways in which the SAM process can influence school leaders and subsequently their schools, teachers, and students. This is a hallmark of program evaluation. Program evaluation requires not only knowing what a program expects to achieve, but also how. Weiss (1995) refers to a theory of change approach to program evaluation, or theory-based evaluation. Program theory "deals with the mechanisms that intervene between the delivery of program services and the occurrence of outcomes of interest. It focuses on participants' response to program service. The mechanism of change is not the program activities per se but the response that the activities generate" (p. 73). In contrast to the SAM process, many evaluations of PD for school leaders do not attend to the conceptualizations and program theories that can explain how program developers and implementers expect the PD experiences to influence leaders, teachers, students, and their schools (Goldring, Preston, & Huff, 2012; Grogan & Andrews, 2002). Thus, we believe an added benefit of an RCT will be to contribute much to the overall knowledge base on program evaluation for school leadership development.

As noted previously, this is not to gainsay the fact that participants face challenges in learning and implementing the SAM process, especially those aspects that go beyond reconciling the calendar, holding the SAM Daily Meeting, and other straightforward process components. These challenges are an area of concern. The hypothesized outcomes of the SAM process rest not only on increasing the *amount of time* on instructional leadership, but also in increasing the *quality* and *specificity* of instructional leadership. These areas of the theory of action are complex, and it is not clear that the SAM process has sufficiently addressed the training, support, and expectation-setting necessary for schools to address them fully. We heard some of these concerns in our case studies and Time Change Coach and Implementation Specialist interviews; participants noted that the success of the program rests with principals knowing how to improve the quality of their instructional leadership behaviors and practices, but they also questioned how principals in the SAM process learn about high-quality instruction or how principals learn how to develop cooperative teams of teachers for productive learning communities. These issues are addressed by some coaches and PD opportunities, but they may not be sufficiently and clearly articulated as part of the role of the coach. Even when present, they may not be of sufficient "dose" or intensity to lead toward desired outcomes. We did, however, hear that a handful of principals were working toward improving the quality of their instructional time use.

It also may be important to consider that most of the schools in our sample were early in their adoption of the SAM process. It is unclear whether this is a reason why few were actively working on improving the quality of time use; some coaches suggest that the first year or two of implementation is focused on the mechanical aspects of the process, while later years allow for a

deeper focus. Thus, it remains unclear whether principals would begin to focus on quality in later years of implementation.

These also are areas where more explicit articulation of the roles of principals, SAMs, and coaches are needed. In addition, much more direct and intensive support and training of coaches is required. And there are implications for who the SAM might be. It is one thing to have a secretary as a SAM who is monitoring a calendar (time); it is another to have a SAM who can really work with the principal on learning and reflecting about providing high-quality feedback to teachers. These are issues that are at the front and center of NSIP's ongoing development and work. NSIP is aware of these needs and complexities, and the theory of action now encompasses instructional time quality.

An additional concern is sustainability. Most schools rely on grant funding. Presumably, the costs of the implementation of the SAM process in an RCT would have to be covered by the funding agency for the entire timespan of the study. This was an area of concern for coaches, who saw sustainability as one of the largest challenges to the SAM process.

Criterion 3: Does the SAM process show evidence of efficacy in changing behaviors or outcomes consistent with its theory of action?

The evidence in support of an RCT on this criterion, changing outcomes, is mixed, dependent on the outcome variable of interest. Our answer is yes if the dependent variable is increased emphasis by principals on allocating time toward the improvement of instruction. This is a finding in the Time/Task Analysis shadowing data, comparing time use from one year to the next, and TimeTrack calendar data, comparing first-year participants to veterans. Principals and district staff clearly articulate this benefit in the case studies, as do Implementation Specialists and Time Change Coaches.

Our methodology for this study did not allow us to directly test other outcomes, beyond self-reports through interviews and surveys. Those self-report data do suggest other perceived and plausible benefits and outcomes that are consistent with the theory of action. For example, 46% of principals indicated that the SAM process helps them improve instruction in the school "a lot." Case studies and interviews provided insights into other reported benefits, such as a positive change in school culture, valuing staff, providing leadership growth opportunities for others in schools, and managing a work/life balance. Respondents across surveys and interviews reported that the SAM process does show efficacy in changing behaviors and outcomes.

In this study, we did not test whether the SAM process changes student achievement. Prior evidence from an evaluation conducted by PSA found little evidence of achievement impacts, although as we noted in the main report, limitations faced by that study suggest that the results are largely inconclusive (Turnbull, White and Arcaira, 2010). Regardless, it is important to underscore that the SAM process as currently implemented is different from what it was at the time of the PSA study. In addition, changes in student achievement were not among the most important factors for why principals decided to participate in the SAM process. Case studies and other interviewees noted that many factors contribute to improvement in student achievement, acknowledging the indirect linkages between principal leadership and student outcomes. However, many believed that the SAM process was likely a contributing factor.

We think that the shift in allocation of time to instructional leadership is sufficiently important and sufficiently documented to do an RCT to determine whether there are, in fact, causal relationships between the SAM process and other important outcomes as articulated in the theory of action beyond the amount of time on instructional leadership. We are less sanguine about documenting impacts on student achievement, given the indirect nature of the relationship between changes in principal leadership and student outcomes and the empirical challenges of detecting those effects.

Criterion 4: Do measures exist (or could they reasonably be developed) to capture elements of the theory of action and hypothesized outcomes?

We believe this criterion is met. A test of the theory of action of the SAM process in an RCT rests with the availability of existing measures to test hypothesized outcomes beyond student achievement. In this case, we are interested in both mediating variables, thought of as near-term effects of an intervention that are required for the intervention to have the desired longer-term effects, as well longer-term outcomes.

In the SAM process theory of action, the most important mediating variable is actual changes in time spent on instructional leadership. There are established methodologies and measures that can be implemented both in the Time/Task Analysis shadowing and TimeTrack calendar data as part of the intervention itself, as well as other measurement approaches, external to the SAM process, that have been implemented in the recent literature, such as end of day logs and detailed shadowing observations (Goldring et al., 2008; Grissom, Loeb, & Master, 2013). Surveys, observations, and interviews can ascertain other measures of implementation fidelity.

We have robust measures for many of the other concepts in the theory of action. For example, changes in quality of instruction can be directly measured through the implementation of reliable and valid observation rubrics, such as the CLASS or Danielson Framework for Teaching (see Goldring et al., 2014). These measures have been implemented in RCTs. School culture, climate, and relational trust can be measured through previously developed surveys that have known psychometric properties (see, for example, extensive survey-based climate measures developed by the Consortium on Chicago School Research). Instructional leadership quality can be measured by the Vanderbilt Assessment of Leadership in Education (Porter et al., 2010).

The challenge will be to develop measures of both the quality and frequency of specific instructional leadership behaviors, such as providing feedback to teachers. While these measures have heretofore not been well conceptualized in the research literature, the seeds of such work are beginning with research groups across the United States. For example, the Center for Educational Leadership at the University of Washington has developed scales and rubrics for ascertaining the quality of feedback provided to teachers in post-observation conferences.

Criterion 5: Is an RCT of the SAM process feasible from a design perspective?

An RCT has three main design goals: internal validity, sufficient precision for detecting effects, and external validity. Here we discuss each of these goals and our assessment of the feasibility of an RCT of the SAM process that meets these objectives. Our general assessment is that, with careful attention to these details, an RCT can be designed that meets these goals. Additional considerations around each of the goals are discussed in the next section.

First, an RCT should have *internal validity.* By that, we mean that it should provide unbiased estimates of the treatment effect—in this case, the effect of participation in the SAM process on an outcome of interest. To provide unbiased estimates, there must be a realistic means for randomly assigning a school to the SAM process (treatment) or business-as-usual (control), and it must be possible to guard against such threats to internal validity as spillover between the treatment and control groups, differential attrition, and instrumentation bias. Our assessment is that such a design is feasible. For reasons we describe in more detail later, we recommend a design that recruits districts to offer the treatment to schools, then randomizes among schools that wish to participate, with schools as the unit of assignment. Only principals in treated schools will obtain access to the various components of the SAM process.

Spillover between participating and non-participating principals within districts is likely to be minimal because the SAM process is a school-level intervention whose main components (TimeTrack calendar, Time/Task Analysis results, support/training from NSIP) cannot easily be transferred, although steps may need to be taken to ensure that Time Change Coaches working with treatment principals do not also provide similar supports to control principals in the district, an issue discussed further in the next section. Because spillover can never be completely controlled, researchers should gather data on potential spillover via surveys from control principals during the study.

To minimize differential attrition between the treatment and control groups, the study should be designed so that a school would remain in the study only if it continues implementing the SAM process or remains an uncontaminated business-as-usual control. That is, a school is not lost when a principal leaves the school, so long as the replacement principal is recruited to use the SAM process. Neither is a school lost from the intervention group when a SAM leaves, so long as the SAM is replaced. Schools would be lost from the study if they refused to participate any longer, either because they no longer wished to implement the SAM process or they no longer wished to be in the business-as-usual control group or to provide data.

Instrumentation bias occurs when data collection methods influence the results in either the treatment or control group. Although researchers must take care to guard against instrumentation bias in designing and collecting survey and interview data, we believe the greatest threat of instrumentation bias in a study of the SAM process rests in the observational study of principal time use. That is, changes in principal instructional time are central to the SAM process theory of action, and the current process uses trained observers who are employees of NSIP to shadow participating principals. The shadowing occurs over full weeks at the baseline and once per school year subsequently to provide an unbiased accounting of changes in principal time investments. Any RCT of the SAM process will likely seek to use these observational data and potentially implement similar observations of control group principals so changes in principal time use can be tested as an effect of the intervention. To ensure that data are collected in a consistent manner, the same observers should be used for both treatment and control principals, with attempts made to keep observers "blind" to the treatment (i.e., they should not be told which principals are treatment or control) in the baseline data collection. In subsequent years, this blindness will not be possible—observers will see treatment principals using the TimeTrack calendar, for example—but observers should be trained to collect data similarly regardless of whether observing a treatment or control principal. In addition, researchers may consider dual coding of some observations or other procedures to guard against this possible threat in

collecting observational data. Researchers also may consider collecting shadowing data independently of NSIP, which could accommodate alternative coding schemes for tracking principal time use.

The second goal of an RCT is adequate *precision* (or statistical power). By precision, we mean whether an RCT can be designed to be sufficiently precise to detect an effect size of educational importance on outcomes of interest. Precision can be thought about in terms of the size of the standard error of the estimate of the treatment effect, or the statistical power for detecting a treatment effect of a given size. The primary question for ensuring sufficient precision is whether an adequate number of schools can be recruited to participate in the RCT. Precision also can be enhanced through adding covariates correlated with the dependent variables and/or blocking on variables correlated with the dependent variables, points we discuss in more detail later.

Third, an RCT should have *external validity.* By external validity, we mean that the results can be generalized from the specific study to a larger set of circumstances (e.g., different populations of subjects, different contexts). Designers of an RCT can improve external validity by recruiting districts to participate from a variety of locations and contexts. Aside from choice of study sites, the technical way to explore external validity is to build a design that allows sufficiently powered tests of interactions with variables that might plausibly alter the implementation or effects of the program. For example, given evidence that elementary and secondary school principals engage differently with the SAM process, one might build a design that allows a test of whether the intervention interacts with level of schooling. If there is no interaction, then the intervention has greater external validity than if there is an interaction, which indicates that the size of the effect depends upon whether you are investigating one level of schooling or another (Porter, 1997).

There are at least two important issues to consider in designing a study that incorporates tests for interactions. The first is which variables it would be desirable to test. In the case of the SAM process, likely candidates include years of principal experience (i.e., novice principals may be affected by participation in the SAM process much differently than principals who have been leading schools for many years), school enrollment size or level (given that principal time use is likely very different in small and large schools), and characteristics of the student population that might be associated with principal time demands (such as average prior achievement levels in the school). Researchers also would potentially seek to test for interactions with indicators of SAM implementation, with the expectation that the impact of the SAM process is higher in schools that implement it more rigorously. The second issue is how the need to test such interactions affects the size of the study. The statistical power for tests of interactions typically is low, suggesting that the design will need to include many more schools to allow for reliable tests. For some kinds of variables, oversampling may be necessary, as in the example of schooling level, since most districts have many more elementary schools than secondary schools.

III. Designing a High-Quality RCT of the SAM Process

Having established from our analysis that an RCT of the SAM process is both warranted and feasible, we next provide further recommendations regarding how an RCT might be implemented to meet its goals of achieving high internal validity, precision, and external validity.

Recruiting and Choosing Participating Districts and Schools

From our current data collection and analysis, two principles appear key in recruiting sites and participants for an RCT of the SAM process. The first is that district context is likely to be an important factor in how principals use their time and in how the SAM process is implemented. The second is that the kind of buy-in from principals that comes with self-selecting into the SAM process is essential to its likelihood of success in a school. To address the first principle, we recommend that each district be treated as a "block" to control for district-specific impacts on time use and SAM process implementation—that is, districts would be recruited first, and then randomization would occur within districts. To address the second principle, we recommend that school participation be voluntary, not required. Districts agreeing to partner in the study would offer the opportunity to participate to principals, and randomization would be conducted within the list of principals who sign up.

Although concerns for external validity suggest that researchers should aim to recruit districts from diverse contexts, including with respect to size and location, we recommend that recruitment should focus on districts of medium to large size. Larger districts are necessary to ensure that there are sufficient numbers of principals who seek to opt into the treatment. Larger districts also are more likely to have processes in place to facilitate the sharing of school, teacher, and—if necessary—student data for the evaluation. While including schools from smaller districts (of the kinds typical of rural areas, for example) is desirable, we suspect that the benefits of recruiting more populous districts for an RCT outweigh potential external validity costs. (We did not study SAM process implementation in small, rural districts in our case studies).

We also recommend that an RCT recruit *new* districts (i.e., districts that do not currently have schools participating in the SAM process). There are at least two reasons for this recommendation. First, new districts will start afresh with no preconceived notions of what the SAM process involves. The fresh start makes possible a common definition of the SAM process intervention and the nature of district involvement. Second, new districts will have more eligible schools and less likelihood of the control schools being contaminated by having former SAM principals at the outset or during the period of intervention.

Another important recruitment issue is what incentives might be necessary to ensure initial and continued participation in the study. Given the substantial ongoing growth of the SAM process into new districts willing to pay for the process and NSIP's services, we anticipate that researchers will not encounter trouble with identifying districts willing to serve as study sites in exchange for free provision of the SAM process to participating principals. Although it may be necessary to budget some payment to school districts for data provision, our assessment is that access to the SAM process at no cost is likely a sufficient incentive for district participation. Similarly, incentives are unlikely to be necessary for treatment principals beyond access to the SAM process for their schools. Control principals, however, may require incentives to ensure that they provide survey and/or interview data each year as the RCT unfolds, given concerns about differential attrition from the study as a threat to internal validity. Unless it is cost-prohibitive, we recommend a delayed treatment approach in which control group principals are promised the SAM process at a later time—perhaps in the third year of implementation—as an incentive to provide data on the business-as-usual condition in initial years of the study.

Choosing Sample Sizes

Sample size considerations for an RCT include both the number of schools required and, given the block design we propose, the number of districts that should be recruited. The number of schools required for an RCT is determined by power calculations. These calculations rely on a number of assumptions, including what statistical power is necessary, what proportion of variance in the dependent variable can reasonably be explained by available covariates, and what the minimum effect size is that the RCT needs to be able to detect. Although actual sample size requirements will vary according to the specific design chosen by researchers, to provide an approximation, we calculated what sample would be necessary for a two-level fixed effects blocked individual random assignment design, where principal time on instruction is the dependent variable[1] and half of schools in a given district (block) would be assigned to treatment and half to the control condition.[2] We made the standard assumption of power of 0.8 and assumed that 30% of the variance in the dependent variable could be explained by the block and principal- or school-level baseline covariates.[3] For the minimum detectable effect size (MDES), in the absence of other information, often a conservative value of 0.2 is chosen. Our analysis of the Time/Task Analysis shadowing data, however, suggests that effects of the SAM process on principal instructional time use may be much higher. Growth from baseline to one year after beginning the SAM process for the sample we analyzed was approximately 0.7 standard deviations. Thus, we conclude that it may be sufficient to design an RCT capable of detecting an effect higher than 0.2 but lower than 0.7, given the non-experimental nature of our Time/Task Analysis sample. For an MDES of 0.45, a sample size of 110 schools (split evenly into treatment and control) would be required. An MDES of 0.35 would require 183 schools.[4]

Potential principal turnover may also need to be considered in choosing the number of schools to include. Turnover can be divided into moving to other principal positions in the district and exiting the principalship in the district, either to take another role in the district or to leave the district altogether. Presumably, for the first type of turnover, the principal can remain in the RCT, so this scenario presents less of a problem. For the second type, however, the principal will leave the study, and if the new principal chooses not to adopt the SAM process, the size of the sample is reduced. Although we do not have information on relative turnover in SAM and non-SAM schools, numbers from the 2011-12 Schools and Staffing Survey suggest that, nationally, approximately 15% of traditional public school principals will either leave the district or the principalship per year. This figure suggests that some compensating increase in initial study samples is warranted and also highlights the importance of considering impacts on principal turnover as a potential outcome for the RCT.

---

[1] Note that other potential outcomes of interest that cluster within schools, such as student achievement, require different power calculations based on different assumptions.

[2] Calculations were performed using PowerUp!, a tool for calculating sample sizes and minimum detectable effect sizes for experimental designs introduced by Dong and Maynard (2013).

[3] Other assumptions in these calculations also are standard, such as an alpha level of 0.05 and two-tailed hypothesis testing.

[4] Increasing the amount of variance in the dependent variable explained by covariates also reduces sample sizes. For example, under the assumption that covariate-explained variance is 50%, an MDES of 0.45 can be achieved with 81 schools, while an MDES of 0.35 requires 132 schools. At the same time, including tests for interactions with potentially important covariates, such as school level, will increase sample sizes above these minimums.

The first consideration for how many districts should be included is how many are needed to achieve the requisite number of schools. If 110 schools are necessary, and the typical district that agreed to serve as a study site could recruit 22 principals to participate, then five districts would be required. If districts could only recruit 11 principals, on average, then 10 districts would be needed. The second consideration is external validity. Results from a study that randomizes within a larger number of districts with differing characteristics (e.g., urban, rural) will be more generalizable than a study that includes a smaller number of districts. The third set of considerations, which must be balanced against the benefits of having more district sites, are the cost and complexity of recruiting, coordinating across, collecting data from, and analyzing data from a larger number of districts, which are important but difficult to quantify.

Defining and Standardizing the Intervention

We assume that NSIP, the current provider of the SAM process, would be the provider of the SAM intervention in an RCT, rather than an independent third party, given that the supports provided to SAM participants by NSIP are an integral component of the SAM process. We make this assumption explicit to underscore that an experiment with NSIP delivering the intervention has external validity that only generalizes to NSIP being the delivery mechanism.

The design of a SAM process RCT must begin with a definition of what is being tested. One broad possibility for this definition is that an RCT will test the SAM process as it exists at the beginning of the RCT. Another is that it will test the SAM process as NSIP chooses to implement it over the course of the RCT. These two definitions will not necessarily be the same, because NSIP continually makes new features or components available to SAM principals. These changes, such as the recent addition to the TimeTrack calendar of modules to track and summarize time spent with specific individual teachers, can be valuable to principals but also present somewhat of a moving target in knowing what the RCT is testing. This would have to be explicitly decided and agreed upon before embarking on an RCT.

Relatedly, the current version of the SAM process allows for a good deal of local adaptation, and an important choice point for the design of a SAM RCT is whether such adaptation is conceived of as part of SAM process implementation or whether an attempt should be made to standardize one or more components of the intervention. An example of standardization would be mandating the number of SAMs in the school or what kinds of positions should be SAMs. We recommend that the SAM intervention include local adaptation and choice rather than a highly prescriptive and standardized design. We believe this will allow for a great degree of sustainability, engagement, and motivation for participation and will lead to the best plausible effects. We believe that local adaptation and choice, as is currently in the field, is part of the definition of the SAM process. However, we also recommend that no significant new process or product developments (such as major enhancements to TimeTrack) be introduced to schools participating in the RCT during the study so that the treatment delivered can be clearly defined.

Delivering the Intervention

The SAM process as currently implemented includes consultation between SAM principals and Time Change Coaches, some of whom have other coaching or supervision roles in local school districts. In an RCT, we recommend that individuals employed as Time Change Coaches work *only* with treatment principals in any capacity. In other words, we recommend against a model in

which SAM process coaching becomes part of the portfolio of work of existing district personnel with other responsibilities that lead them to work also with control group principals. Coaching work with both treatment and control principals is a primary potential avenue for spillover of some aspects of the intervention, one of the main threats to internal validity of a SAM process RCT.

Collecting Data on Mediating Variables

A thorough examination of the SAM process via an RCT requires collection of data on mediator variables, or intermediate outcomes in the SAM process theory of action. First, we anticipate that researchers will want to collect measures of principals' time investments to assess how the SAM process changes their allocations across categories. Two sources of such data are the TimeTrack calendar and the Time/Task Analysis shadowing. Because TimeTrack calendar data are unavailable at the baseline, cannot be gathered for control group principals, and rely on principal fidelity to the treatment for accuracy, we recommend that measures of principal time allocation focus on in-person shadowing via the Time/Task Analysis process or other shadowing implemented independently by the research team. Weeklong shadowing data should be collected at the baseline for both treatment and control principals and ideally also would be collected at multiple points over the study years to measure learning associated with the treatment and seasonality in time use.

Second, we recommend that researchers collect measures of *quality of instructional time,* which has become more of the focus of the SAM process in recent years. Such measures likely will require input from the teachers who are the subject of principals' instructional investments via yearly surveys and completion of a principal assessment tool that can capture this dimension of principal work, such as the Vanderbilt Assessment of Leadership in Education. These measures should be collected in both treatment and control schools.

Surveys of SAMs, teachers, and other school personnel also can capture other potentially important mediating variables from the theory of action, including relational trust, distributed responsibilities, and teachers' attitudes toward their instructional practice. Principal surveys can be used to capture principal work/life balance measures and measures of principal reflective practice.

Measuring Impacts on Instructional Practice and Student Achievement and Implications for Study Duration and Cost

The SAM process theory of action shows that changes in principals' instructional time investments should ultimately lead to changes in teachers' instructional practices and, subsequently, student achievement. We recommend that substantial attention be paid to how to capture these two "final" outcomes of the SAM process. Unfortunately, both present substantial difficulties. Measuring changes in teacher instructional practices in a valid and reliable way may require in-person or video observation (and subsequent coding) of teacher instruction on a large scale in both treatment and control schools, which comes with large resource costs. Measuring impacts on student achievement present difficulties in this context because we have little guidance on how long to expect it will take for changes in principal instructional time use—mediated through a large number of other variables—to be reflected in changes in student achievement scores. We anticipate that a five-year window may be necessary to see such effects

to give principals time to change their behavior, teachers to change their behavior in response, and students' achievement to respond to an improvement in instruction.

This window has direct consequences for the duration of the study. Fortunately, descriptive evidence from both Time/Task Analysis and TimeTrack calendar data suggests that changes to principal time use as a result of participation in the SAM process likely happens quickly and may be evident even after the first year of the intervention (although getting the SAM process up and running proves a challenge in many schools, according to our survey and interview data). Given uncertainty about the time required to see impacts on instructional practice or student achievement, we recommend that schools be funded for five years of intervention, which suggests a seven-year study: one to implement the RCT design, five years of intervention, followed by one year to complete analysis and write-up of results. We note that an RCT that did not seek to assess impacts on student achievement presumably could be reduced to 2–3 years of intervention, with accompanying reductions in total study length.

NSIP provides SAM process services to schools for $12,900 in the first year. For 110 schools—one estimate provided in the section on sample sizes—this is a cost of $1,419,000 (assuming the control schools would also eventually receive SAM process services under a delayed-treatment design). Subsequent years would be lower cost; NSIP reduces its fees each year that a school participates. The total cost of delivering the SAM process in the context of an RCT would depend on these reductions and how many years control schools would be provided access.

Securing Access to Data

Access to data on SAM process implementation, fidelity, and mediating and final outcomes is essential to success of an RCT. Collection of survey, interview, and administrative data for both treatment and control schools for the duration of the study likely will be required in any RCT design. In the present study, securing access to relevant data was hampered by the relatively low numbers of principals who provided consent to NSIP to release their data or to allow NSIP to provide the research team with their contact information for surveys or interviews. Based on this experience, we recommend that an evaluation study team enter into data sharing agreements with districts and schools directly, not through NSIP, to provide these data to researchers, including consenting to allow NSIP to provide internal data on participating schools to researchers (such as TimeTrack data), as a condition of participation at the outset of the study. Furthermore, we recommend that structures be put into place whereby the research team works with school districts directly to obtain data or solicit school personnel for participation, rather than working through NSIP. The research team should work directly with NSIP to obtain internal data on principals, SAMs, and others as necessary—for example, internal TimeTrack calendar data, which is likely to be valuable in capturing implementation fidelity—but again, researchers should obtain permissions from schools themselves directly as a condition of the study that will allow NSIP to provide those data.

Roles for NSIP

As provider of the SAM process (the treatment), NSIP would play an integral role in an RCT. We suggest that the research study team recruit more districts than needed initially that meet research study criteria (i.e., number of schools for random assignment, not a district with current SAMs), and then that the study team would work with NSIP. NSIP's role would begin with

assistance recruiting from the pool of potential districts to participate, given the organization's frequent contact with potential new district partners, but the research study team would have the final decision about district participation. Given the importance of principal buy-in, we anticipate that NSIP would also play a key role in presenting and "selling" the SAM process to potential principals in study sites to gain their voluntary participation, a step that NSIP's director says the organization takes with potential new SAM principals already. Once schools are selected to participate in the SAM process, we recommend that NSIP's relationship to the research team be "at arm's length." NSIP should provide SAM process support to treatment principals just as it would with any principal, without special consideration for a principal's role in the study. To this end, it would be preferable for treatment principals to be "blind" to NSIP (i.e., NSIP would not know which principals are participating in the RCT), but this step is infeasible for a number of reasons, including the desirability for researchers to obtain TimeTrack calendar and other internal NSIP data on participants for purposes of studying implementation. We do recommend, however, that study-related communication with treatment principals come from the research team or through the district partner, to keep the study and implementation of the SAM process as separate as possible.

References

Berman, P., & McLaughlin, M. W. (1978). *Federal programs supporting educational change, vol. VIII: Implementing and sustaining innovations*. Santa Monica, CA.

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24–67.

Goldring, E., & Preston, C. & Huff, J. (2012). Conceptualizing and evaluating professional development for school leaders. *Planning and Change, 43(3), 1-13.*

Grissom, J. A., Loeb, S., & Master, B. (2013). Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educational Researcher, 42*(8), 433–444.

Grogan, M., & Andrews, R. (2002). Defining preparation and professional development for the future. *Educational Administration Quarterly*, *38*(2), 233–256.

Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), *New Approaches to Evaluating Community Initiatives Volume 1: Concepts, Methods, and Contexts* (pp. 65–94). Washington, DC: The Aspen Institute.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research*, *78*(1), 33–84.

Porter, A. C. (1997). Comparative experiments in educational research. In R. Jaeger (Ed.), *Complementary methods for research in education (second edition)* (pp. 523–551). Washington, DC: American Educational Research Association. (Revised chapter and reading from 1988 edition.)

Porter, A. C., Polikoff, M. S., Goldring, E., Murphy, J., Elliott, S. N., & May, H. (2010). Investigating the validity and reliability of the Vanderbilt Assessment of Leadership in Education. *Elementary School Journal, 111,* 282–313.

Turnbull, B.J. Arcaira, E. & Sinclair, B. (2011). *Implementation of the National SAM Innovation Project: A Comparison of Project Designs.* Washington, DC: Policy Studies Associates, Inc.

Turnbull, B. J., Haslam, M. B., Arcaira, E. R., Riley, D. L., Sinclair, B., & Coleman, S. (2009). *Evaluation of the school administrator manager project*. Washington, DC: Policy Studies Associates, Inc.

Turnbull, B.J. White, R.N., & Arcaira, E.R. (2010). *Achievement Trends in Schools With School Administration Managers.* Washington, DC: Policy Studies Associates, Inc.

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. P. Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), *New Approaches to Evaluating Community Initiatives Volume 1: Concepts, Methods, and Contexts* (pp. 65–94). Washington, DC: The Aspen Institute.

ABOUT THE WALLACE FOUNDATION

The Wallace Foundation is a national philanthropy that seeks to improve education and enrichment for disadvantaged children and foster the vitality of the arts for everyone. The foundation works with partners to develop credible, practical insights that can help solve important, public problems.

Wallace has five major initiatives under way:

- **School leadership**: Strengthening education leadership to improve student achievement.
- **After-school**: Helping cities make good after-school programs available to many more children, including strengthening the financial management capacity of after-school providers.
- **Building audiences for the arts**: Developing effective approaches for expanding audiences so that many more people might enjoy the benefits of the arts.
- **Arts education**: Expanding arts learning opportunities for children and teens. Summer and expanded learning time: Better understanding the impact of high-quality summer learning programs on disadvantaged children, and how to enrich and expand the school day.
- **Summer and expanded learning time**: Better understanding the impact of high-quality summer learning programs on disadvantaged children, and how to enrich and expand the school day.

VANDERBILT PEABODY COLLEGE

The Wallace Foundation®