EVALUATING A LARGE-SCALE, MULTI-INSTITUTION PROJECT: CHALLENGES FACED AND LESSONS LEARNED

Erica Slate Young
Appalachian State University
slateer@appstate.edu

Bryanne Peterson
Virginia Tech
bryanne@vt.edu

Sarah Schott
Duke University
schott@math.duke.edu

Jack Bookman
Duke University
bookman@math.duke.edu

https://doi.org/10.25891/5e14-nf34

ABSTRACT
SUMMIT-P consists of nine participating institutions working toward common goals but from unique perspectives. Evaluating such a large-scale project with diverse stakeholders has presented challenges. For one, evaluation on this scale necessitates a team effort rather than a single evaluator. Communication is key among the evaluators as well as among the project players at large. Participation and reliable, timely feedback from participants are perhaps the most important issues while also posing some of our greatest challenges. We present strategies we developed to counteract these challenges. In particular, we discuss the development of an assessment tracking system used to not only monitor responses but to also promote an increase in on-time responses. We conclude with a discussion of some lessons learned about evaluating large-scale, multi-site projects to share with other evaluators and PIs alike.

KEYWORDS
multi-site evaluation, educational research

As educational research projects evolve in the 21st century, evaluation of these projects is evolving as well. Technology allows for patterns and changes to be explored at a greater scale and at distance, which has led to collaborative opportunities to explore change and growth across multiple sites. One such example of this type of project is A National Consortium for Synergistic Undergraduate Mathematics via Multi-institutional Interdisciplinary Teaching Partnerships (SUMMIT-P).

Background of SUMMIT-P

SUMMIT-P is a curriculum and faculty development project spread across multiple institutions and designed to implement the recommendations from the MAA Curriculum Foundations (CF) Project (Ganter & Barker, 2004). The project is funded by a grant from the National Science Foundation (NSF). The member institutions of SUMMIT-P form a diverse consortium in that they vary by size and type as well as geographic location. Each of nine institutions has formed interdisciplinary teams to organize discussions with local faculty from one or more partner disciplines about how best to implement changes in the lower division undergraduate mathematics courses to reflect the needs of students in those partner disciplines. In addition, these local interdisciplinary teams are expected to:

- organize discussions with local faculty in mathematics and the partner disciplines to make use of insights about interdisciplinary collaboration from the CF reports,
- organize frequent internal project team meetings to discuss course content, development/progress of work, and necessary alterations to the work plan,
- appoint one member of the institution's key personnel to be responsible for working with the central evaluation team to collect institutional data while ensuring that all partners within the institution provide necessary information, including information about faculty members' perceptions of the impact of the intervention on students' attitudes, skills, and vocational interests,
- participate as an interdisciplinary team in regular communications and meetings with the consortium wide project team,
- visit and host several site visits with commonly aligned institutions within the collaborative, and
- contribute to the national impact of the project by reporting on their work in publications and national meetings.

The project aims to create an enduring network of faculty and programs within and across institutions to share experiences and ideas for successfully creating functional interdisciplinary partnerships.

Important Elements of Program Evaluation Relevant to SUMMIT-P

Every NSF-sponsored curriculum reform project is required to have a program evaluation component. In order to support program evaluators in their work, the NSF has produced a useful and clear handbook for conducting program evaluations, *The User-friendly Guide to Program Evaluation* (Frechtling, 2010). In this guide, evaluation is defined as follows:

> A comprehensive definition, as presented by the Joint Committee on Standards for Educational Evaluation (1994), holds that evaluation is "systematic investigation of the

> worth or merit of an object." This definition centers on the goal of using evaluation for a purpose. (p.3)

Frechtling (2010) continues to summarize three main purposes of evaluation: (a) to produce information that could help to improve a particular project, (b) to document what has been done on the project, and (c) to potentially gain new insights that were not expected. "What are frequently called 'unanticipated consequences' of a program can be among the most useful outcomes of the assessment enterprise" (Frechtling, 2010, p.3).

Another essential element in conducting a program evaluation is the communication between the evaluators and the stakeholders. Alkin et al., (2006) state the importance of this element by saying, "communication is a part of all program evaluation activities. Indeed, it is probably not an exaggeration to say that evaluation without communication would not be possible," (p.385). In a large-scale project such as SUMMIT-P, the importance of communication is magnified. In our situation, the stakeholders also serve as what Frechtling (2010) defines as "key informants" (p. 71). Key informants are those who have, "unique skills or professional background related to the issue/intervention being evaluated, [are] knowledgeable about the project participants, or [have] access to other information of interest to the evaluator," (p.71). Therefore, communication within this project must be a two-way street. Not only are we, the evaluators, responsible for communicating with the stakeholders, the stakeholders, as key informants, must be in communication with us.

Exploring Large-Scale Program Evaluation

In this paper, we will discuss the process of conducting a large-scale program evaluation, focusing on not just our methodology but also the successes and difficulties we have encountered; in particular, we will present our solution to the communication challenges that occur in a large-scale, multi-site evaluation. In conducting the program evaluation of large-scale projects such as SUMMIT-P, having a diverse project with respect to the types of institutions involved is both a strength and a challenge. While this diversity allows us to examine how the project evolves in many different settings, it has also been the source of many challenges we have faced. The responsibility for the evaluation team is to examine the progress being made towards the SUMMIT-P project goals, as stated below:

- Implement major recommendations from the MAA Curriculum Foundations (CF) Project for the purpose of broadening participation in, and institutional capacity for, STEM learning, especially relative to teaching and learning in undergraduate mathematics courses;
- Foster a network of faculty and programs in order to promote community and institutional transformation, through shared experiences and ideas for successfully creating functional interdisciplinary partnerships within and across institutions;
- Change the undergraduate mathematics curriculum in ways that support improved STEM learning for all students while building the STEM workforce of tomorrow; and,
- Monitor how various aspects of the CF recommendations are being implemented at participating institutions while measuring the impact on faculty and students.

Evaluating progress made toward these ambitious goals would be a challenge to evaluate on even a small scale; considering the magnitude of this project, this task is monumental. A one-size-fits-all evaluation model is not adequate for a project of this scope. In the sections that follow, we

will discuss some of the particular challenges faced along with a few strategies we have implemented in our efforts to overcome these challenges.

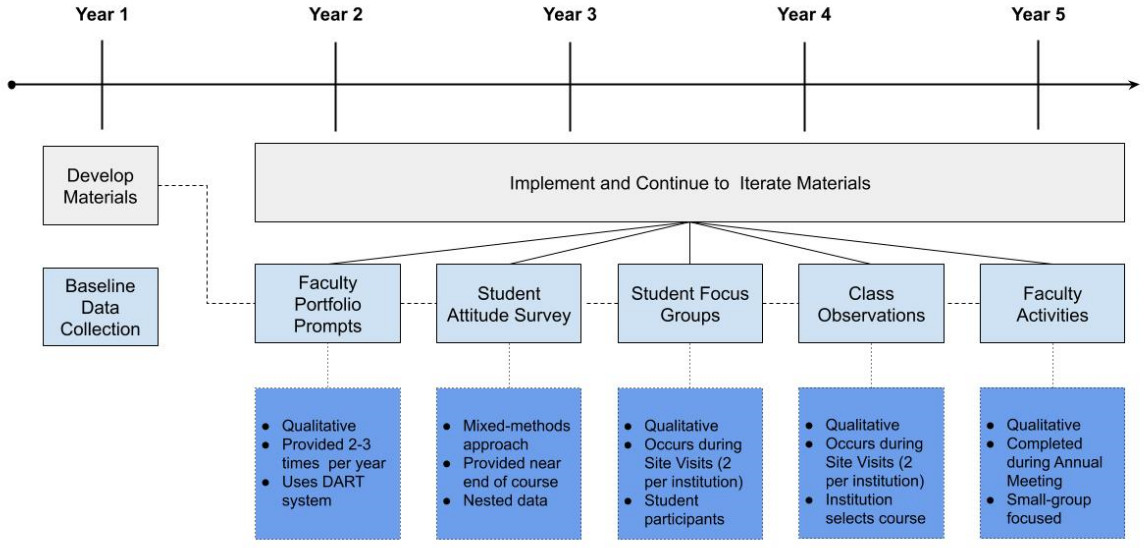
As mentioned above, this project is a collaboration among nine institutions with diverse backgrounds and populations. While all the institutions have a common goal, each institution is implementing its own model of change within the SUMMIT-P framework. This requires the evaluation team to utilize Multisite Evaluation (MSE) methods to conduct the evaluation and research. As stated in Straw and Herrell (2002), "two factors differentiate MSEs from other evaluation activities: the involvement of multiple sites and the conduct of a cross-site evaluation activity" (p. 5). Our evaluation activities are aimed at examining the program as a whole as well as the implementation at each of the sites.

The SUMMIT-P Evaluation Model

The fact that SUMMIT-P is being enacted through a large consortium of institutions creates many logistical challenges for the program evaluation. "The larger the number of sites, the more important are standards for data collection, quality control, and data submission," (Rog, 2010, p. 100). From the start, the plan was for the evaluation to be conducted by a team rather than a single external evaluator. On a project of this scale, the effort of many minds is better than the single perspective of an individual.

Best evaluation practices dictate the use of a mixed-methods design for our research and evaluation efforts due to the nature of the objectives we are assessing (Frechtling & Sharp, 2002). In order to measure the impact the program is having on student outcomes, we are collecting survey data from students enrolled in affected courses, survey data from faculty involved in teaching those courses, and other qualitative measures. Of particular interest to us is the examination of change in this context, one of the main themes of the research questions.

Figure 1
Evaluation Timeline



In addition to baseline survey data from both students and faculty, we used the SUMMIT-P site visits to further triangulate our data collection efforts (see Figure 1). As part of

the project, each institution will host two site visits: the first in year two or three of the project, the second in year four or five. At site visits, a team of people travel to the host institution to observe their efforts and progress made toward the SUMMIT-P goals (Piercey & Segal, 2020). This team is composed of Primary Investigators (PIs) and co-PIs from one or two other SUMMIT-P institutions, a representative of the Project Management Team, and a member of the Evaluation Team. These site visits provide us with the opportunity to observe classes where lessons developed as part of the SUMMIT-P project are being taught (SUMMIT-P lessons), participate in both formal and informal conversations with the various stakeholders at the institution, and also conduct focus-group sessions with students whose classes are utilizing SUMMIT-P lessons. We are therefore able to examine the evolution of the project at each site firsthand.

Furthermore, we spend time during the annual SUMMIT-P face-to-face meeting, held in conjunction with the Joint Mathematics Meetings, to conduct focus group sessions with PIs and co-PIs. It should be noted that, although some data is being collected from students enrolled in the courses, our primary subjects are the faculty (i.e., the 39 PIs and co-PIs) involved in the project.

One of our richest sources of data is the Evaluation Portfolio we created in which we list various prompts for the faculty to respond to several times over the course of the year. The first prompt was designed for participants to provide baseline data regarding their prior teaching experiences and teaching philosophies. Participants wrote a “teaching autobiography” wherein they described their teaching experiences over the years, focusing on elements such as: their first teaching experiences, how their philosophy of teaching has changed over time, what teaching methods they employ, how their beliefs about student learning have changed over time (if at all), and what they find most challenging and most rewarding about teaching. Because our aim was to collect baseline data to help us develop a deeper understanding of the participants we are studying, this was a relatively lengthy writing task. The subsequent prompts have been designed for shorter responses and therefore require much less time for participants to complete. Here are examples of some of the other prompts participants have responded to:

- When all is said and done, what would convince you that your project was successful? In other words, how would you define "success" within the context of your specific situation?
- In general, when you are looking for ways to change your teaching, where do your new ideas come from? Tell us about the kinds of sources from which you primarily draw new ideas.
- Recall a recent conversation or interaction with a colleague or a student related to the work of SUMMIT-P. For example, this could be from a discussion in class or office hours, in a planning meeting with colleagues, a conversation with a dean or other administrators, etc. In your response briefly tell us about what was said (just enough to give us the main idea of the interaction/conversation). What did you learn from this interaction?

Crafting prompts that allow participants to share how the project has evolved from their perspective is essential to our research model. This is one of the main benefits of using an Evaluation Portfolio to collect data. Rather than create a list of predefined questions at the beginning of the project to be asked at regular intervals, we designed the prompts to address specific questions that are based on what is relevant to the project at the time. For example, the prompt which asked participants to define “success” within their particular context was based on

a discussion between the PIs about the progress being made at their respective institutions. We decided that we needed to hear from all of the participants regarding their definitions of "success" to help us evaluate the progress at each individual institution. We are also interested to see if the definition changes for any of the individuals over the course of the project. We will use the Evaluation Portfolio to ask a follow-up question in the final year.

Since our primary emphasis is on faculty growth, our main artifact for analysis is the Evaluation Portfolio collected from the participating faculty who, as stated above, serve as our *key informants*. We are studying the responses using qualitative content analysis (Mayring, 2000). We will be triangulating our analysis with the data collected through the faculty survey that is being conducted during the site visits and the annual face-to-face meetings. Additionally, in order to develop a more complete picture of the changes taking place within the various institutions, we are collecting student data through surveys, class observations, and focus group sessions conducted during site visits. We will analyze the student surveys using a factor analysis to determine trends in their responses. We will implement a grounded theory approach to coding and analyzing the information collected through class observations and focus group responses.

In typical survey research, as with the student survey we are administering, a 30–50% response rate is considered acceptable. Because of the relatively small number of faculty participants and due to the qualitative nature of our work, we need nearly a 100% response rate to the faculty prompts in order to generate valid results. Considering that the faculty who are being surveyed are also working together on this project and are being partially supported by the grant, we believe this is a reasonable expectation. In order to show growth, we need consistent participation from PIs and co-PIs at all stages of the project: the beginning, the middle, and the end. In other words, we need responses to reach a "critical mass" in order for them to be representative.

Challenges Faced

MSEs inherently come with a set of challenges and our situation is no different. There are a large number of sites, and, by design, each one is unique. We have dealt with unanticipated events such as a change in PI at some institutions. For example, one institution withdrew from the project after two years because the PI accepted a position at a different institution. Also, across all of the project PIs, there is a broad range of prior experience with large-scale funded projects. In qualitative (or quantitative) studies where one is trying to document growth, collecting high quality baseline data is important. This can be a big challenge to project evaluation because of the inevitable unforeseen circumstances, such as changes in project personnel.

We also must consider the significance of individual personalities to program evaluation. Understanding the personalities of the individuals involved is a crucial factor in creating a functional team dynamic. An individual's personality and motivations play a role in project success. This phenomenon has been documented in the literature: "[m]uch also depends on such social factors as political and intellectual alliances, friendships, and institutional loyalties" (Bell, 1998; Godfried, 1999 as cited in Leff & Mulkern, 2002, p. 90).

While being able to examine how the project evolves in many different settings is beneficial, this is simultaneously the source of challenges. With 10 sites, 39 PIs and co-PIs, additional instructors of courses using SUMMIT-P lessons, and over 4,000 students involved, gathering data and tracking responses requires additional oversight and the use of advanced

metrics. Moreover, based on the wide range of experience and understanding of project evaluation among the individuals involved in the project, it has been important to spend time explaining and discussing the purposes of program evaluation with the entire group. This has been essential for everyone to understand the significance of evaluation to the SUMMIT-P project. As described in the evaluation model above, the PIs and co-PIs are prompted to submit data including student attitude surveys and responses to evaluation portfolio prompts regularly. Dealing with stakeholders who are non-compliant or demonstrate low levels of participation is a challenge as there has been little by way of repercussions outside of the inconvenience of being asked repeatedly. Closely related to this challenge is the difficulty of finding an effective communication platform. It became apparent early on in the project that email alone was simply not sufficient. Information needs to be communicated among all parties, and, in addition to the evaluation team, PIs need to be able to keep track of the data collection requirements and deadlines. Project evaluators and participants need to know what has been submitted and what is still outstanding.

During the first two years of the project, we attempted to address the problem with email in a number of ways. We looked carefully at semester schedules and deliberately set data submission deadlines to avoid the busiest times of the academic calendar. Initially, we thought that sending email reminders would be sufficient to increase response rates. First, we sent a reminder directly to those individuals who had not yet responded. If that did not produce a response, a second reminder was sent to the participant and the local institution PI was also copied. If necessary, the third email reminder was sent to the participant, the PI, and the lead project director. While multiple reminders did increase the response rates somewhat, they were an inefficient use of the evaluators' time, especially considering that we did not reach our desired response rates.

At the second annual SUMMIT-P face-to-face meeting the evaluation team gave a presentation discussing the importance of the evaluation efforts. We reviewed the project research questions and the goals for the evaluation. We shared the response rates for the various data collection measures and explained how the evaluation depends on timely responses and that it is possible to submit a response that is "too late" to be useful. Our hope was that by educating the participants on the issues underlying the evaluation efforts the response rates would improve. In general, this was not effective. Our response rates improved slightly but were still not at acceptable levels.

While the evaluation team was actively working on a solution to improve response rates, we sought input from the Project Management Team regarding this challenge. Based on their input, we voluntarily participated in a Descriptive Consultancy Protocol during a virtual PI meeting, described in Hobson-Hargraves et al. (2020).

The purpose of a Descriptive Consultancy Protocol is to find a solution to a dilemma during a discussion session with a neutral, skilled facilitator. The person (or persons) with the dilemma poses the problem; then the group, in this instance the PIs at the virtual meeting, restates how they interpret the dilemma. After this, the group brainstorms solutions to the dilemma. "The justification behind this protocol is that framing and reframing a complicated problem is valuable for moving towards a focused solution" (McDonnough & Henschel, 2015, p.147).

We were looking for PI input on what parts of the system would be important to them, and we also wanted to give them some agency over the solution. By giving them some ownership of the process, our hope was that they would be more invested in data collection

success. A solution was necessary because, up until that point, response rates were hovering around 25%. Personalized email follow-ups were taking up to six person-hours, per prompt, of evaluation team time, and yet the highest response rate achieved was 70%; this was not a sustainable system. The Descriptive Consultancy Protocol allowed us to work together with the PIs in order to fine-tune a potential solution to the response rate problem. This solution, which we have named the Digital Automated Response Tracking (DART) system, is described below.

Our Solution

We needed a system that would be straightforward for the participants to use and that would collect the data in an organized way. Additionally, we needed to monitor progress simply and efficiently, preferably in real time, and have a way to share that monitoring responsibility with institutional PIs. With Google Forms as a basis for portfolio response submissions, a tracking system was developed in Google Sheets. Using a single, stable web link, participants are able to access all current and previous prompts. They select their name, institution, and which prompt they intend to answer. The prompt question is used for skip-logic branching to lead the participant to the appropriate question set where they submit responses in a "long answer" field. This element alone streamlined the process significantly as participants were previously sending responses via email which the evaluation team then had to compile into a central repository.

The real power of the DART system is the built-in tracking feature on the back end. Within the Google Suite, when a participant completes the response to a particular prompt, that response is automatically logged on the Form Responses tab by Google. A master tracking spreadsheet with the participant names listed as the rows and the status of each prompt (i.e., complete or incomplete) in the columns was created using the Google Sheet linked to the Google Form. This tracking list is a worksheet (also called a tab) that draws the data directly from the Form Responses tab in real-time, using an array formula. That addition is automatically noted by the array formulas on the tracking sheet and the appropriate cell changes from incomplete to complete for that prompt. A corresponding conditional formatting change occurs (i.e., red to green) as well. See Figure 2 for a sample of the tracking form.

Figure 2

Sample of the DART System Tracking Sheet

Institution (Institution names removed to protect privacy)	Person (Individual names removed to protect privacy)	Position	Institution Response Rate	Prompt 1- Spring 2017	Prompt 2- Fall 2017	Prompt 3 - Spring 2018	Prompt 4- Summer 2018	Prompt 5- September 2018	Prompt 6- Spring 2019
Institution A	[Participant 1]	pi	100%	Submitted	Submitted	Submitted	Submitted	Submitted	Submitted
Institution A	[Participant 2]	co		Submitted	Submitted	Submitted	Submitted	Submitted	Submitted
Institution A	[Participant 3]	co		Submitted	Submitted	Submitted	Submitted	Submitted	Submitted
Institution A	[Participant 4]	co		Submitted	Submitted	Submitted	Submitted	Submitted	Submitted
Institution B	[Participant 5]	pi	94%	Submitted	Submitted	Submitted	Submitted	Submitted	Submitted
Institution B	[Participant 6]	co		Submitted	Submitted	Incomplete	Submitted	Submitted	Submitted
Institution B	[Participant 7]	co		Submitted	Submitted	Submitted	Submitted	Submitted	Submitted
Institution C	[Participant 8]	pi	63%	Submitted	Submitted	Submitted	Incomplete	Submitted	Submitted
Institution C	[Participant 9]	co		Submitted	Submitted	Submitted	Incomplete	Submitted	Submitted
Institution C	[Participant 10]	co		Submitted	Incomplete	Incomplete	Submitted	Submitted	Incomplete
Institution C	[Participant 11]	co		Submitted	Incomplete	Incomplete	Incomplete	Incomplete	Submitted

Figure 3
Sample of DART Institution-level Tracking Sheet

Institution C

Person (Individual names removed to protect privacy)	Position	Prompt 1- Spring 2017	Prompt 2- Fall 2017	Prompt 3 - Spring 2018	Prompt 4- Summer 2018	Prompt 5- September 2018	Prompt 6- Spring 2019
[Participant 8]	pi	Submitted	Submitted	Submitted	Incomplete	Submitted	Submitted
[Participant 9]	co	Submitted	Submitted	Submitted	Incomplete	Submitted	Submitted
[Participant 10]	co	Submitted	Incomplete	Incomplete	Submitted	Submitted	Incomplete
[Participant 11]	co	Submitted	Incomplete	Incomplete	Incomplete	Incomplete	Submitted

Team response rate: 63%

From this master tracking document, there are separate tabs for each institution that import the data via an index function. Separate Google Sheets for each institution were then created that use an IMPORTRANGE function to link each school's data to the new sheet; this allows for real-time updates to be visible in the new Google Sheets without access to other institution submission records. Institutional PIs were then granted "view only" shared rights to the sheet for their institution, allowing PIs to view the information for their institution without having the ability to modify it. Each of the lead PIs has access to his or her institution's sheet, therefore allowing them to track in real-time who has submitted responses and to which prompts. See Figure 3 for a sample.

Reflections on the Effectiveness of the New System

Thus far, the DART system has been an effective tool. The first submission collected via email, prior to implementing the DART system, had a response rate of 36% before reminder emails and 71% after two rounds of reminder emails; the most recent prompt had an 87% submission rate with no individual email reminders sent out. Response rates are much improved and, equally importantly, the time required to monitor the submissions has drastically decreased. Because the lead PI at each institution can see their own response data (including who has and who has not yet responded) they are in a better position to encourage and monitor participation, removing this burden from the evaluation team. Another benefit of the DART system is that it allows a "one-stop shop," so to speak, for participants to catch up on prompts they have not yet completed. They do not need to dig through their email inbox to find and respond to overdue prompts; instead, they are able to complete any of the evaluation portfolio prompts within DART using a single web link.

Another advantage of this system is that it houses all responses in one spreadsheet, allowing the evaluation team to easily read through responses to a particular prompt. Responses can be exported to another Google Sheet or Google Doc effectively and efficiently by using Google Suite add-ons which allow us to annotate documents with our comments and conduct a content analysis.

Many of the challenges we are facing in this MSE are par for the course in examining a complex system such as the SUMMIT-P consortium project. Analyzing data from multiple institutions will always be a challenge for evaluators, especially when there exists significant

variation from one institution to another. There will always be challenges associated with unexpected events, such as a change of the PI at an institution, and there is no way to control for the "human element" inherent in the personalities, expertise, and priorities of the people involved in the project. However, we are pleased to have found a solution to one significant challenge faced in this MSE through the implementation of the DART system.

Lessons Learned

In this section, we will summarize the lessons we have learned thus far in conducting this MSE. We have organized our thoughts into two sections—advice to other evaluators and advice to future PIs.

Advice to Other Evaluators

One of the most important pieces of advice we can pass along to other evaluators, in particular those conducting MSEs, is to always focus on the people. Whether they are your subjects of analysis, those facilitating the project, or, as in our case, serving in both roles, the individual personalities, the group dynamics, and the institutional cultures will all play a large role in your work. It is important to take all of these elements into account when planning for and then carrying out the evaluation activities.

Also, there is never a "good time" for participants. Time is a precious commodity for everyone, especially in academia; everyone is busy. It is important to set clear expectations for participation up front but also to be prepared to be flexible as needed. Participants need to be willing to make time for evaluation activities, and evaluators need to be willing to adjust evaluation plans when warranted. We have found clear, effective, and efficient communication to be helpful here, a lesson we learned at times the hard way. By providing a scaffolded data collection system like DART from the beginning, investigators could set clear expectations while also providing scaffolding for successful data collection rates. It is recommended that a response system like DART be included in the data collection plan of a proposal to ensure its use from the beginning. By anticipating and preparing for the complications that come with an MSE in the planning process, evaluators can focus more on the evaluation content.

Advice to Future PIs

Communication plays a key role in the advice we offer to both future evaluators and PIs. One important role of the PI regarding the evaluation activities involved in a large-scale project is to establish a communication plan at the start of the project. The role of a local PI in a multi-site consortium, such as SUMMIT-P, is different from that of a PI of a single-site project. In a multi-site grant, the local PIs play the role of "middle management." They communicate with not only the local project participants but also with the project leadership and other PIs as well. Anticipate that email alone might not be sufficient.

Additionally, ensure you have adequately budgeted time and resources for the evaluation at both the institutional and consortium levels. In a large-scale MSE, the evaluation cannot be an afterthought; it must be an integrated part of the project. There should be clear expectations for how things will be done as well as the required levels of participation in evaluation activities. This is crucial if there are qualitative aspects of the MSE, which there very likely will be. A

single evaluator will likely not be adequate; an evaluation team is probably needed. In addition to the budget for the project-level evaluators, consider adding a percentage to each local budget to support a PI or co-PI who will be responsible for overseeing project evaluation and research efforts at each site, who would then report to the central evaluation team.

Our final piece of advice is to be flexible. Expect and be prepared for unanticipated events. Have a clearly laid out plan, but know that just as projects evolve as time goes on, evaluation plans must evolve, too.

Conclusions

Project evaluation conducted by external evaluators is more than just a required element in grant work; it is a crucial piece of the project being implemented. External evaluators provide a birds-eye view of the project that few others involved in the project are able to see. We are entrenched in the work being done. We are able to see all of the various parts of the project and how they are (or are not) working together. In a large-scale consortium like SUMMIT-P, this birds-eye view is even more important. We hope that the lessons we are learning as part of our work on this project will help others who may be involved in MSEs—both participants and other evaluators. Each project comes with its own unique set of challenges. Likewise, each project needs a customized evaluation plan that can be responsive to those unique challenges; however, the lessons we have shared in this paper can be broadly applied.

Our parting thoughts are as follows: projects are composed of many elements—the "products", the people doing the work, the data being collected, the students being taught, etc. However, in all of this, the importance of the personalities of the people involved cannot be overstated. With the right groups of committed, dedicated people working together, great things can get done.

Overall, the main lesson we have learned from this project can best be encapsulated in a quote from Holley (1982), "Those being evaluated often feel threatened by the evaluation. Evaluators need to accept the behaviors of evaluation subjects. They must be patient, persistent, and persuasive" (p.1). It all comes down to this: in the end, it is always about the humans.

Acknowledgment

This paper was developed in part through the project *Collaborative Research: A National Consortium for Synergistic Undergraduate Mathematics via Multi-institutional Interdisciplinary Teaching Partnerships* (SUMMIT-P, www.summit-p.com) with support from the National Science Foundation, EHR/IUSE Lead Awards 1625771, 1822451, 1942808. The opinions expressed here are those solely of the authors and do not reflect the opinions of the funding agency.

References

Alkin, M. C., Christie, C. A., & Rose, M. (2006). Communicating evaluation. In I.F. Shaw, J.C. Greene., M.M. Mark (Eds.), *The SAGE handbook of evaluation* (pp. 385 – 403). Thousand Oaks, CA: Sage Publications Ltd.

Frechtling, J. (Ed). (2010). *User-friendly handbook for project evaluation*. Arlington, VA: National Science Foundation. Retrieved from https://www.purdue.edu/research/docs/pdf/2010NSFuser-friendlyhandbookforprojectevaluation.pdf

Frechtling, J. & Sharp, L. (Eds.). (1997). *User-friendly handbook for mixed method evaluations*. Arlington, VA: National Science Foundation. Retrieved from https://www.nsf.gov/pubs/1997/nsf97153/start.htm

Ganter, S. and W. Barker (Eds.). (2004). *The curriculum foundations project: Voices of the partner disciplines.* Washington, DC: Mathematical Association of America.

Hobson, R., Hofrenning, S., Bowers, J., Beisiegel, M., Piercey, V., & Young, E. (2020). Structured engagement for a multi-institutional collaborative to share best practices and tackle challenges. *The Journal of Mathematics and Science: Collaborative Explorations*, *16*, 43 – 57. https://doi.org/10.25891/vf7a-ps53

Holley, F. M. (1982). *Foundations of reporting: Or Bartlett's guide to evaluation communication* (ED216309). ERIC. https://files.eric.ed.gov/fulltext/ED216039.pdf

Leff, H. S., & Mulkern, V. (2002). Lessons learned about science and participation from multisite evaluations. *New Directions for Evaluation, 2002*(94), 89 – 100.

Mayring, P. (2000). Qualitative content analysis. *Forum: Qualitative Social Research*, *1*(2), Art. 20. http://dx.doi.org/10.17169/fqs-1.2.1089

McDonnough, J.T. & Henschel, M.M. (2015). Professional learning community-based induction: Creating support for new teachers of science. In J. A. Luft & S. L. Dubois (Eds.), *Newly Hired Teachers of Science* (pp. 145 – 153). Sense Publishers.

Piercey, V., Segal, R., Filappas, A., Chen, T., Kone, S., Hobson, R., Bookman, J., Hearn, J., Pike, D., & Williams, K. (2020). Using site visits to strengthen collaboration. *The Journal of Mathematics and Science: Collaborative Explorations, 16*, 27 – 42. https://doi.org/10.25891/wpxj-gg40

Rog, D. (2010). Ensuring rigor in multisite evaluations. In J. Frechtling (Ed.), *The 2010 User-Friendly Handbook for Project Evaluation* (pp. 97 – 111). National Science Foundation.

Straw, R. B. & Herrell, J. M. (2002). A framework for understanding and improving multisite evaluations. *New Directions for Evaluation, 2002*(94), 5 – 16.

Sukovic, S. (2017). Evaluation & research. Part 2: Similarities & differences. *Health Education and Training*. Retrieved from https://blog.heti.nsw.gov.au/2017/10/25/evaluation-research-part-2-similarities-differences/