12-2019

Promoting STEM Teacher Reflection through Self-Evaluation

Steven W. Mumford
University of New Orleans, swmumfor@uno.edu

Kathryn Newcomer
George Washington University

Promoting STEM Teacher Reflection through Self-Evaluation

Steven W. Mumford & Kathryn Newcomer

Author Information

Steven W. Mumford, PhD, Assistant Professor, Department of Political Science, University of New Orleans, New Orleans, LA, steven.mumford@gmail.com.

Kathryn Newcomer, PhD, Professor, Trachtenberg School of Public Policy & Public Administration, George Washington University, Washington, DC, newcomer@gwu.edu

Correspondence concerning this article should be addressed to Steven Mumford, Department of Political Science, 256 Milneburg Hall, University of New Orleans, 2000 Lakeshore Drive, New Orleans, LA 70148, Tel: 206-484-4661, Fax: 504-280-3838.

Acknowledgments

Funding for this research was provided by the National Science Foundation's Robert Noyce Scholarship Program under award number 1240041. The authors wish to thank our program partners and participants who contributed greatly to this research, as well as our graduate assistants who provided critical background research and support.

Abstract

Evaluators face serious challenges assessing the impact of teaching practice on student outcomes, such as for assessing the impact of grant-funded teacher training programs. To overcome these barriers, the evaluators of an NSF-funded teacher scholarship program elected to coach teachers on conducting their own evaluations of student outcomes. To facilitate this process, we collaboratively developed templates and guidance for teachers' measurement and reporting of student growth, provided training and technical assistance as teachers prepared student outcome data, and guided teachers in reflecting on results to improve their practice. The evaluators then aggregated fellows' student outcome reports to assess the program's overall impact. We discuss challenges that prompted the approach, explain how we improved it over time, and reflect on strengths and shortcomings. Beyond contributing to summative program evaluation, this model for teacher self-evaluation enhanced participating teachers' capacity for self-reflection and ability to continuously improve their teaching practice beyond the fellowship.

Keywords: assessment, capacity building, self-evaluation, STEM education, preK-12 education

Promoting STEM Teacher Reflection through Self-Evaluation

Continuously improving, "data-driven schools" have long served as a vision for well-meaning educational reformers, with teachers positioned at the forefront of student data collection and use (Noyce, Perda, & Traver, 2000). The theory of change goes something like this: Through systematic evaluation processes (Isore, 2009), teachers are encouraged, or even required, to apply data to intentional, critical reflection on the results and underlying assumptions of their teaching, thereby improving their practice and students' learning. Rather than treating teaching as a "mystery" grounded primarily or even solely in personal experience, "critically reflective teachers" are expected to subject their experience to data-driven critical analysis and develop explicit rationales guiding their teaching practice (Brookfield, 1995).

The promotion of teacher reflection connects closely with evaluation scholars' growing interest in intentionally developing stakeholders' evaluative thinking skills through capacity-building and process use (Vo & Archibald, 2018; Buckley, Archibald, Hargraves, & Trochim, 2015). Given the high stakes of student learning outcomes, particularly in STEM fields (Saxton et al. 2014), and the potential for evaluation systems to either mitigate or exacerbate inequities depending on how they are implemented (Carden, 2017; Vo & Archibald, 2018; Donaldson, Woulfin, LeChasseur, & Cobb, 2016), it seems that professional evaluators would have much to offer to advance teacher evaluation and reflective practice. However, studies of efforts to develop teachers' reflective practice are in their infancy (Alemdar, Cappelli, Criswell, & Rushton, 2018; Saxton et al., 2014; Henderson, Beach, & Finkelstein, 2011), as are explicit theoretical connections to the evaluation literature (Archibald, Neubauer, & Brookfield, 2018).

In the United States, the reality of public schools does not match the lofty rhetoric of "data-driven decision making." Recent educational reforms across levels of government have

emphasized summative use of student test data in teacher evaluation systems (Croft, Roberts, & Stenhouse, 2015; Economic Policy Institute, 2010), potentially at the expense of sufficient time and resources devoted to formative approaches focused on teacher improvement (Donaldson et al., 2016; Isore, 2009). Indeed, teachers face cultural (e.g., personal metrics for success), political (e.g., mistrust of top-down policies), and technical barriers (e.g., time and skills) to the use of accountability-driven standardized test data for improving their practice, and may instead rely on anecdotal information, experience, and intuition as guides (Ingram, Seashore Louis, & Schroeder, 2004).

Teachers would be better served by data they personally perceive as credible and relevant, which suggests they should actively participate in design of the evaluation system (Smagorinsky, 2014; Donaldson, 2012). In turn, locally-driven teacher evaluation systems that teachers find meaningful for feedback and improvement may enhance their intrinsic motivation, autonomy, and self-confidence (Firestone, 2014; Ross & Bruce, 2007), and students' learning outcomes (Ozogul & Sullivan, 2007).

How might evaluation professionals help engage teachers in a meaningful process of systematic self-evaluation (Isore, 2009)? Research suggests that external facilitators can promote teacher self-improvement by establishing and supporting a cycle of individualized, data-informed feedback and reflection, involving the teacher's completion of a self-designed action research project (Hardre et al., 2014; Henderson et al., 2011; Ross & Bruce, 2007). Teacher training in self-evaluation is critical to the success of such an approach through instigating and guiding the overall process, and specifically promoting the perceived objectivity of data, modeling and coaching the delivery and use of feedback, and providing instruction and practice in technical skills (Ozogul & Sullivan, 2007).

Indeed, established outcomes measures and methodologies are generally lacking for evaluating teacher performance, including in the STEM fields, despite policymakers and funders' growing interest in science and math teacher development programs in recent years (Hardre et al.; Saxton et al., 2014; Alemdar et al., 2018). As a result, teachers need guidance in designing credible research projects (Noyce et al., 2000), like tailored advice for implementing a pre-assessment (Guskey, 2018). Unfortunately, teachers may rarely receive this guidance or support from principals and specialized instructional coaches or team leaders (Donaldson et al., 2016), even when called for in district policy (National Research Council, 2015).

The authors were contracted to evaluate a new, NSF-funded teacher scholarship program designed to recruit and develop middle and high school science and math teachers in Washington, DC. We were expected to secure student performance data to assess the effectiveness of the intervention, towards the ideal of a rigorous evaluation design focused on student learning outcomes (e.g., Institute of Evaluation Sciences, 2013). However, we were concerned by the questionable credibility of student performance data as evidence of teachers' abilities (Economic Policy Institute, 2010; Isore, 2009), not to mention resource constraints and other practical considerations common in small-scale evaluations of teacher training programs (Hardre et al., 2014; Alemdar et al., 2018). Therefore, we developed an evaluation capacity building approach in which we explicitly targeted the program participants' evaluative thinking skills (Vo & Archibald, 2018), and thus their motivation and ability to engage in critical reflection (Brookfield, 1995).

In this article, we share our process for evaluating outcomes of the teacher training program—both in terms of student outcomes and teachers' continuous improvement—and discuss strengths and weaknesses of the approach. We begin with a brief description of the

teacher scholarship program. Next, we walk through our self-evaluation approach step by step, including the processes of supporting participating teaching fellows in the design, implementation, analysis, reporting, and further refinement of their own evaluations of student learning and growth. Finally, we reflect on the strengths and challenges of the approach, and how it might be useful for educational evaluators and other practitioners going forward, particularly for evaluating STEM teacher training programs.

Overview of the Teacher Training Program

We evaluated an NSF-funded Robert Noyce Teacher Scholarship program (e.g., Alemdar et al., 2018), implemented by a local university's graduate school of education in collaboration with a private, non-profit organization already providing similar programming in the area. The goal of the program was to increase the number of highly qualified math and science teachers in high-need schools in Washington, DC. The program included two sequential cohorts of up to 16 teaching fellows each (eight math, eight science) who participated over a six-year period, from 2012 to 2018. A total of 30 teaching fellows, roughly evenly divided between math and science fields and cohorts, started the program over its first two years, and 26 (87%) completed it. **Figure 1** displays the theory of change we developed with program staff during the grant application stage to depict the intent of the program.

[Figure 1 Here]

Program Elements

The teacher scholarship program (hereafter "program") began in late 2012 with the recruitment of a diverse group of participating "fellows" with professional – but not always teaching – backgrounds in math and science. The fellows participated in an intensive one-year, 30-credit master's degree in teaching to prepare for licensure, a practicum and student teaching

semester with support from faculty and a "cooperating teacher" at the school placement, and workshops and assistance with navigating the application process at local schools.

Through the first four years of teaching after completion of the degree and licensure, the program provided additional mentoring and support from faculty and "master teacher" mentors, in the form of multiple observations per year and individualized feedback. The program also supplemented the fellows' salaries during their four-year post-degree teaching commitment in Washington, DC, public and charter schools. The program provided professional development opportunities throughout its duration. Lastly, fellows were contractually required to participate actively in the grant evaluation process throughout its duration as a condition of their award.

Rationale for Our Evaluation Approach

The grant-mandated external program evaluation was designed to promote accountability, provide feedback and data for learning and program improvement, and share results and lessons learned with the broader field. Towards these ends, the evaluation provided continuous formative feedback to program staff, collected through biannual online surveys, self-assessments, and/or phone interviews with participating fellows, as well as review of documentation of program implementation, fellows' academic performance, and fellows' ability to obtain and retain a teaching position in a high-needs school in Washington, DC. Data summaries were shared with program staff regularly to promote real-time learning and use of findings for program improvement; results were also compiled in interim and final reports.

This approach was in line with guidance provided elsewhere for evaluating STEM teaching training programs (Hardre et al., 2014; Saxton et al., 2014), and specifically a Noyce Scholarship program, typically characterized by evolving program elements, small samples, and nuanced outcomes, all of which benefit from more in-depth qualitative methods (Alemdar et al.,

2018). Not surprisingly, more rigorous and quantitative designs–requiring large samples and substantial resources to isolate teacher's impact on student learning (e.g., Institute of Evaluation Sciences, 2013)–were quickly deemed unfeasible, and as mentioned earlier, were not expected to provide formative feedback most useful for teacher self-reflection and improvement (Donaldson et al., 2016; Isore, 2009; Ingram et al., 2004).

Therefore, we placed a large emphasis on helping the fellows develop evaluative thinking skills that they could incorporate in their teaching during their careers (Vo & Archibald, 2018). We felt it was essential that fellows take ownership over their evaluation planning, implementation, and interpretation of findings, so that they could tailor their approach to their own learning style and preferences for credible data (Smagorinsky, 2014). Nonetheless, in addition to formative questions about fellow satisfaction and development, the funding agency (NSF) and program staff were interested in a key summative question: "To what extent have K–12 learners in classes taught by Fellows improved their STEM skills?"

This inquiry forced us to confront the myriad challenges involved when trying to assess the effect of any one program on student outcomes. Particular challenges included: the diversity of subjects, grade levels, and curricula taught by fellows; the variety of school contexts and cultures (including both public schools and public charter schools); changes in standardized testing by District of Columbia Public Schools (DCPS) over the course of the program (including replacement of math exams and development of a new series of science exams); and general skepticism around teacher "value-added" analyses (Economic Policy Institute, 2010; Isore, 2009). These challenges and others emerged from the context for evaluation–both in terms of the local policy context for teacher evaluation, and fellow's unique school contexts and evaluation needs–which we describe next.

Local Policy Context

As a result of multiple concerns, value-added models (VAMs) of teacher effectiveness are generally discouraged (Economic Policy Institute, 2010), at least as the sole or even primary assessment of teacher effectiveness (Isore, 2009). Despite these concerns, as the teacher training program launched, DCPS began implementing a new teacher evaluation system called IMPACT, which relies heavily on calculating teachers' statistical value-add to student standardized test score growth, at least in math subjects where these data are most readily available (National Research Council, 2015). The system is high-stakes, with implications for teacher promotion, pay, or dismissal, but there is also a coaching component, based on observations and tailored feedback from instructional coaches. Unfortunately, the instructional coach positions are often under-resourced and under-utilized, and principals are not always positioned to provide formative feedback for a variety of reasons, including lack of time and power imbalances with teachers (Donaldson et al., 2016).

For teachers of subjects in which standardized tests are not available (such as science, although district-wide assessments are in the works), DCPS teacher evaluations take into account "teacher-designed assessments" of student learning, similar to the individualized action research approach we implemented; however, this process appears to lack guidance, quality control, and documentation (National Research Council, 2015). In brief, the DCPS IMPACT teacher evaluation system is focused on accountability, rendering summative judgments of teacher effectiveness through student test data, rather than on providing useful formative feedback (Isore, 2009). Outside of DCPS, the complex networks of DC public charter schools are free to design their own evaluation systems, which may or may not resemble IMPACT.

Fellows' Evaluation Needs

To better inform our understand of fellows' particular contexts and needs for evaluation, during their first year of full-time teaching, we interviewed fellows about their school contexts and assessment strategies, perceptions of credibility of student outcomes data, logistics of implementing school assessments, and desired student outcomes. Twenty-one active fellows completed a half-hour phone interview on their school's evaluation context, for a response rate of 70 percent. Key themes from interviews are summarized below, providing a snapshot of the diversity and breadth of fellows' school contexts and evaluation systems. Note that many fellows changed schools at least once during the required four-year teaching assignment in DC schools; thus, these themes are illustrative but do not comprehensively describe fellows' school contexts throughout the program's duration.

School contexts. The 21 fellows interviewed taught a wide variety of courses in both math and science fields, from 6th to 12th grade, and at multiple schools about evenly divided between DCPS and a variety of DC public charter school providers. Math courses included pre-algebra, algebra, geometry, pre-calculus, calculus, and statistics. Science courses included biology, chemistry, physics, physical science, environmental science, microbiology, and health. Many fellows also worked with students missing pre-requisites and entering the school year below grade level, and/or with English Language Learners (ELL) or students needing educational supports (SPED). Typical summative approaches to teacher evaluation, such as VAMs, have the potential to negatively bias ratings of teachers who predominately serve ELL, SPED, and low-income students (Economy Policy Institute, 2010; Donaldson et al., 2016).

Targeted student outcomes. Fellows were asked what types of student outcomes they were most interested in assessing, and the desired student outcomes varied greatly. Some of the most common outcomes fellows prioritized included students' content mastery, but also

confidence and enjoyment of the subject matter, problem-solving and perseverance, critical thinking skills, good scholarly habits, student engagement and participation, English language acquisition (for English Language Learners), and graduation and pursuit of STEM (for high school seniors). Fellows, by and large, thus did not view narrow assessments of student learning based on mastery of content as sufficient for self-evaluation, a finding with empirical support in the literature (Smagorinsky, 2014; Isore, 2009; Ingram et al., 2004). Rather, research suggests that effective formative evaluation requires teachers to be closely involved in setting their own goals and criteria for success (Firestone, 2014; Donaldson, 2012).

Challenges to quantitative evaluation. There were several challenges in the evaluation design and data collection processes. Most fellows did not already administer comparable pre- and post-tests with their students, except for those who benefitted from standardized school-wide approaches like those found in International Baccalaureate (IB) curricula, or "iReady" adaptive assessments administered multiple times per school year. Fellows who lacked pre-assessment data were encouraged to add this element into their evaluation plans to reap the methodological and formative benefits (Guskey, 2018), and/or to compare test scores to adopted curricular standards or similar classes in the department, school, or district/charter network.

Even in programs that did have pre-/post-exams, evaluators still had to deal with other issues of credibility that complicated the analysis, discussed later in this article. We thus saw a need for fellows to investigate additional quantitative indicators that might help mitigate validity threats associated with student test scores (Isore, 2009), including pass rates and measures of participation and attendance, as well as qualitative evidence of student growth. Fellows also identified a desire for more focused training and support in basic statistical analysis.

Initial uses of student data. Despite these challenges, we found in our initial interviews that fellows had already begun using formative data to assess and anticipate student needs, revisit content they struggled with, refine exams (e.g., through item and test analysis to remove low-performing questions), and to a limited extent, report results to school administrators. Ongoing feedback from informal checks for understanding with students, such as "exit tickets," warm-ups, and quizzes, was deemed as especially useful by fellows. We hoped to build on fellow's inherent interest in real-time, continuous feedback, while also helping them collect high-quality data appropriate for summative evaluation of the overall teacher training program, and useful for deeper reflection on their teaching practice and its impact on student outcomes.

Steps in the Self-Evaluation Process

In light of the myriad challenges we faced collecting and aggregating student test data across fellows to assess program impact, as well as the opportunities and fellows' desires for formative feedback, we developed an innovative approach to teacher self-evaluation. Fellows were asked to: develop their own individualized evaluation plans for collecting credible data on student content mastery and other outcomes (e.g., student engagement); implement their evaluation plans during the school year to collect student outcomes data (e.g., change in pre-/post-test scores); and report results to us. Our role as evaluators was to act as partners and coaches, gathering feedback on every step of the process, providing templates and guidance, offering workshops and ongoing technical assistance, and aggregating results across the program. We describe the three steps in this self-evaluation process in detail below, and display them in **Figure 2**. We completed the full cycle twice with fellows, over two years.

[Figure 2 Here]

Step 1: Fellows Developed Evaluation Plans

Our first step in the process was to guide fellows in the development of individualized evaluation plans, which would help them collect customized evidence of teacher effectiveness, reflection, and learning. Because opportunities for teachers to self-evaluate and learn are most powerful and efficient when embedded in pre-existing school procedures (Donaldson et al., 2016), as well as the wide variety of school and classroom contexts in which fellows worked, a single, prescribed evaluation methodology was impractical.

Therefore, we developed an open-ended evaluation plan template and shared it with fellows for feedback and approval at an in-person meeting. See **Table 1** for the guiding questions from the final evaluation plan template fellows were responsible for completing during the summer after their first year of teaching. We then reviewed fellows' school-specific evaluation plans and suggested improvements. This process allowed us to tailor evaluations to each fellows' unique needs and context.

[Table 1 Here]

As the guiding questions from the evaluation plan template suggest, fellows were asked to include the following elements in their plans: a pre/post measure of student learning, focused on mastery of subject-specific content; additional quantitative indicators, such as measures of effort or participation; and additional qualitative evidence of student learning, such as sample student work and survey responses. In addition, fellows were asked to discuss at length how they would store and statistically analyze quantitative data, often through an online gradebook, and how and when they intended to reflect on the results, adjust their teaching approach, and share findings with other external parties such as school administration.

Fellows' evaluation plan contents. Almost all active fellows at the time (23 of 26, or 88%) submitted an evaluation plan to the evaluators as required, and a few worked with

evaluators to refine their initial draft based on our feedback. About half of fellows who submitted a plan described a straightforward pre-/post- student assessment process, generally using a "preview" pre-assessment (Guskey, 2018), followed by a comprehensive summative exam repeating similar or identical content. Pre-tests were almost always planned for October, and post-tests in May or later, with some fellows incorporating midterm assessments in December or January.

Many of these assessments were created by fellows themselves, especially for science courses, while others made use of standardized math assessments, practice and actual AP exams, and school-mandated comprehensive exams. Fellow-created exams were often adapted from standardized exams, but also included one test of pre-requisite math skills for advanced physics, and one assessment of evidence-based writing for ELL math.

The remaining fellows either employed school-wide adaptive assessments like iReady that provide grade-level equivalencies for evaluating student growth (especially in math courses), or employed a series of unit-specific pre-/post-exams that could be compared and averaged across the school year. Fellows also tended to include additional quantitative evidence of student attendance, effort, and grades, and qualitative evidence such as student work samples and projects, student survey ratings and open-ended feedback, and student testimonials or stories. These plans served as useful guides and tools for holding fellows accountable for self-evaluation.

Evaluation plan modifications. After their first year of implementation, fellows restarted the self-evaluation cycle at Step 1 and modified their plans, such as by adapting methods to a new school, courses, or grade levels, incorporating a pre- and/or midterm assessment, and replacing teacher-designed exams with validated, off-the-shelf instruments. Interestingly, over time, fellows (especially math) increased their reliance on standardized test

data, in part because DCPS adopted a new district-wide math assessment called the Measures or Academic Progress (MAP). In addition, fellows moved away from reporting student surveys and, to a lesser extent, grades and failure rates, in favor of reporting data on practice and homework completion tracked and monitored over the course of the school year.

To better incentivize student effort on the pre-assessment, several fellows planned to share individual growth scores with students and celebrate their accomplishments. This modification often accompanied the addition of a midterm exam, allowing fellows to calculate change scores in the middle of the school year, rather than waiting for the end when they were busy with other concerns, chiefly end-of-year summative testing. The midterm assessment was also thought to give fellows a mid-year opportunity to reflect on their results and make improvements to their teaching that would directly benefit their students. Although fellows took these and other steps to provide increasingly high-quality data, the process was perhaps most useful for internal reflection and learning on the fellows' part, as opposed to summative evaluation of the program's impact.

Step 2: Fellows Implemented Evaluation Plans

The fellows were then responsible for implementing their evaluation plans over the course of the school year, with support from evaluators. We completed two full cycles of implementation over two years. Each year the evaluators administered a mid-school year survey around the holiday break to remind fellows to collect pre-test data, check in on their progress, and get feedback on any challenges they faced. Fellows sometimes had to adjust their evaluation plan mid-year (in part because of last-minute changes in school policies and assessment procedures), and the survey provided an opportunity to proactively address these concerns and enhance the quality of fellows' ongoing data collection efforts. Fellows generally reported not

needing additional support and technical assistance from the evaluators mid-year, although their lack of time and competing priorities over the course of the school year may have discouraged them from reaching out for help and focusing adequate attention on self-evaluation.

Threats to validity and reliability of student data. As they gained experience implementing their plans, fellows identified multiple potential threats to the validity of their student outcomes assessments, communicated to us in reports of results, and during a focus group conducted as part of an end-of-year meeting and workshop with fellows. Fellows identified threats and potential biases related to selection, mortality, instrumentation, testing, and their interactions; in other words, all the threats plaguing pre-/post-test designs (Shadish, Cook, & Campbell, 2001). In addition, there were statistical artifacts related to small sample sizes and idiosyncratic test scoring procedures.

Fellows reported their most prevalent challenges to the validity of evaluation results as a large number of students missing data and/or entering the school year already below grade level, and therefore unprepared to master new material. Fellows reported pre-tests suffered from low student completion and effort because they covered unfamiliar material, were not graded or otherwise incentivized, and not accommodated for students' special needs (unlike for most post-tests, especially if administered school-wide). Further, class rosters did not stabilize until as late as mid-semester for some fellows, after pre-data were collected.

Fellows noted additional challenges with their student populations and school contexts that impeded rigorous testing. Students were generally described as suffering from test fatigue, and having high prevalence of ELL and SPED needs. ELL students were hampered by standardized tests and other platforms targeted at native English speakers with limited, if any, translation support. Similarly, SPED students were hampered by lack of testing

accommodations, described above. Lastly, several fellows started the school year in a new school, or with a new curriculum, forcing them to adapt lessons and assessment strategies.

Reporting guidance. To bring the second step in the process to a close for each school year, the evaluators convened fellows in late spring to gather reflections on implementation of their evaluation plans, share and refine an additional template for reporting results, and provide brief introductory training in statistical analysis (e.g., reporting descriptive statistics, calculating tests of significance in Excel). See **Table 2** for the original template's guiding questions.

[Table 2 Here]

The evaluators later revised the reporting template for the second year of implementation, replacing open-ended questions with checkboxes where possible based on the first year's overall responses, in an effort to streamline reports. See the **Supplementary File** for the full final template, including table shells for reporting results. During summers, the evaluators provided remote feedback to fellows as they completed their reports, including guidance for organizing and analyzing de-identified student outcomes data.

Step 3: Fellows Reported Results

The third and last phase in the evaluation cycle required fellows to report their results to the program evaluators. Like the process for co-designing fellows' initial evaluation plans, we reviewed a first draft of each fellows' report and suggested changes. After the first year of full implementation of the self-evaluation process, 20 of 26 active fellows at the time (77%) submitted complete student outcomes reports and related de-identified data files. In the second year of implementation, 21 fellows (81%) submitted complete reports, including one additional math fellow. A similar small group of about five fellows failed to complete baseline interviews, evaluation plans, check-in surveys, and final reports for both years, despite repeated reminders

and requests from the evaluators. The evaluators could only relay the list of non-compliant fellows to program staff in an effort to enforce the evaluation mandate in fellows' contracts.

Student outcomes results. Where possible, we aggregated results by the assessment strategies employed and used additional data to add context and individual examples. Based on self-reported estimates, fellows taught a total of about 1,400-plus students in the first year of evaluation implementation, and almost 1,950 students in the second year, grouped into about four to six total classes per fellow. In the second year (the only year for which sample size information was collected), fellows reported outcomes data from a total of about 1,700 students, or approximately 87 percent of all students taught, far higher than expected from fellows' concerns for missing data mentioned above. In sum, each year the majority of fellows provided evidence of statistically and practically significant growth in average change scores from pre to post from their students, using a variety of different assessment strategies.

For the roughly half of fellows (mostly math) who reported student outcomes data from standardized tests (including MAP, iReady, and AP exams), results such as changes in student scores were provided in a format unique to the exam, and in most cases, raw data were not available for significance testing. These fellows shared average growth in students' scores or grade-level equivalency over the course of the school year, whether this growth was practically significant (i.e., did it approximate expectations for a grade-level's worth of learning), and the percentage of students who met or exceeded growth goals or demonstrated "proficiency" (defined as a passing score based on school policy). A few fellows who taught AP classes also shared their students' distribution of scores on the actual exam. These standardized data were not easily interpreted or combined, and potentially not sensitive enough to capture subtle growth,

adding credence to the notion that accountability-driven testing data may not be useful formatively (Donaldson et al., 2016; Isore, 2009; Ingram et al., 2004).

For the remaining half of fellows (mostly science) who used self-created or school exams, results were most often provided in terms of average percentage-point increases in student exam scores, as well as the percentage of students who achieved "proficiency" on a post-exam. With some guidance from evaluators, these fellows were able to use de-identified raw data to test for statistically significant student growth. By the end of the second year, two-thirds of these fellows (i.e., excluding those who provided standardized test data discussed above) reported statistically significant growth in average student exam scores for all classes and student groupings, an increase from just over half of fellows after the first year. The remaining fellows reported non-significant or mixed results by class. In addition to grouping students by class, fellows also analyzed student growth by grade, level (e.g., honors), and ELL or SPED status. One fellow found that students selected for an intensive "math lab" showed the highest average growth of any sub-grouping.

Additional indicators. As described above, fellows could select from a variety of different quantitative and qualitative indicators to add context and nuance to their student outcomes results. The supplemental data provided either reinforced the positive assessment results discussed above by cross-validating post exam and growth scores with other indicators of student effort and content mastery, or they added context and nuance to less positive results by couching them in terms of challenges with student attendance, effort, and behavior. These supplemental analyses are summarized below.

Fellows who reported grades and pass rates noted that students who failed were almost always chronically absent or tardy and/or demonstrated behavioral problems and lack of effort in

class. Two fellows found that growth in assessment scores was positively correlated with student grades. For attendance and work completion, one fellow who administered a "growth mindset" survey to students (Dweck, 2007) found a statistically significant average increase in scores, but little correlation between homework completion and change in growth mindset score. While one fellow reported a strong positive correlation between practice attempts on an online study program and final exam scores, another found no correlation between attempts and growth.

Student surveys were employed much less often than expected based on initial evaluation plans, and they were often removed from the plans after the first year. Two fellows found that as student's satisfaction with the learning environment decreased, based on student survey ratings, student self-reports of their own learning increased. Another shared that the highest-achieving students were able to set reasonable goals on student surveys. Surveys were also useful for determining what lessons and activities students preferred, and which tools best helped them.

Finally, fellows were given the opportunity to summarize qualitative data. Two fellows discussed changing their class format to a "flipped" classroom in which students watch video lectures at home and work on problems in class (McLaughlin et al., 2014). One of them immediately observed increases in student test scores after the change. Other fellows described adapting educational projects like a "shopping spree," and applying learning tools like interactive notebooks and low-tech white boards. Almost all fellows shared observations about challenges in their school context and administration that impeded success, most often administrative requirements that sapped their time, energy, and resources away from formative improvement.

As the results above indicate, individual fellows found that grades, attendance, and work completion and practice data showed potential positive correlations with growth and final exam scores, although these results were mixed. Fellows also added context and nuance to their pre-

/post-exam results. Most importantly, the additional data allowed fellows to explore questions of personal interest to their pedagogy and conduct mini-experiments on their teaching strategies, thereby assessing results and adjusting strategies for the current or future years.

Fellow reflections on evaluation use. Fellows utilized data to reflect on their teaching practices and introduce formative improvements where appropriate, helping them tailor teaching strategies to students in the following ways: identifying student knowledge gaps and misconceptions for re-teaching and tutoring (including in areas like critical thinking and study skills); grouping and differentiating students within lessons; recommending students for course level changes (either in the current or subsequent year); and sharing scores with parents to get them engaged in students' success.

Fellows also shared data with collaborating teachers to strategize about lessons, and fellows increasingly shared growth on the interim or final exam with students to promote a sense of accomplishment. Two fellows shared students' results via bar graphs and used the activity as a statistics lesson and celebration, with one handing out awards for most improved scores. In summary, fellows most often used evaluative data for ongoing reflection and formative improvement of their scaffolding, lessons, and student grouping and differentiation.

Fellow perspectives on evaluation. By and large, despite the significant challenge to their time and resources, fellows were actively engaged in the self-evaluation process and diligent in their data collection, analysis, and reporting, as indicated in the high response rates and quality of reports we received. Throughout the grant evaluation, fellows self-reported increases in their teaching confidence and classroom management skills, and the self-evaluation process may have contributed to those outcomes. In their final reports at the end of the second full year of self-evaluation, all 21 fellows signaled their intention to continue implementing a

self-evaluation and self-reflection process beyond the duration of the program. Indeed, fellows who left the area and even the teaching profession after the program pledged to adapt self-evaluation strategies to their new contexts and careers.

For the most part, fellows planned to sustain and enhance their current evaluation approaches in future years. Fellows pledged to continue refining their assessments and finding opportunities to share growth with students to incentivize and reward effort and growth. Science fellows shared that they might benefit from adoption of a DC-wide standardized science exam in the coming years. However, one math fellow noted that standardized MAP scores did not prove useful for formative feedback, a finding echoed in the literature and discussed throughout this article. Many fellows reported that end of year results in general are not timely enough to help them adapt their teaching during the school year, especially if they planned to change schools or course loads, and some fellows lost access to data at the end of the school year.

Instead, fellows hoped to incorporate more interim exams and continuous analysis and reporting into their evaluation plans, in order to promote regular reflection and course corrections, perhaps in collaboration with mentors and school administrators. Many fellows followed this process naturally, such as by keeping a journal and regularly reviewing "exit tickets," but desired a more formalized and systematic process, especially to avoid bias and document results to outside parties. Nonetheless, fellows reported lacking the time and intention to compile and analyze data during the course of the busy school year. While we were not able to overcome all of these myriad challenges, the fellowship evaluation process appeared to establish a routine for fellows to incorporate into their teaching practice going forward.

Lessons Learned

The process of grant-driven, evaluator-supported teacher self-evaluation described above was by no means perfect, but it provided a methodology for compiling evidence of student outcomes across a diverse group of fellows and contexts, with the benefit of promoting fellows' evaluation capacity, self-reflection, and improvement (Archibald & Vo, 2018). Indeed, fellow reports and interviews offer evidence that they enhanced their capacity for evaluative thinking and critical reflection on their teaching, although more research would certainly be needed in this emerging area of scholarship (Archibald et al., 2018). Our process, supports, and templates provided at each step might offer guidance to evaluators in K-12 STEM education and other fields about how to navigate the significant challenges facing formative evaluation in an accountability-driven era.

We distilled this six-year experience down to five lessons learned about promoting teacher reflection through self-evaluation, some of which we were able to implement, and others we recommend for future implementation and study:

1. **Teachers need an impetus and consistent support to engage in meaningful self-evaluation**. Research shows that teachers benefit from external motivation to participate in evaluation and self-reflection, at least initially (Ross & Bruce, 2007; Noyce et al., 2000), and our contractually-required program evaluation served that purpose. However, impetus is not enough. Teachers also needed support to target meaningful and realistic student outcomes and criteria for success (Firestone, 2014; Donaldson, 2012), identify and collect locally-relevant measures of student learning from multiple sources (Saxton et al., 2014; Isore, 2009; Ingram et al., 2004), and analyze, report, and apply findings (Donaldson et al., 2016; Ozogul & Sullivan, 2007; Noyce et al.). In other words, they needed training, technical assistance, and guidance through every step of evaluation, in order to promote the perceived validity,

reliability, and utility of results (Smagorinsky, 2014; Henderson et al., 2011). The evaluators' role thus shifted from external researcher to partner and accountability coach.

2. **Existing school procedures tend to prioritize high-stakes, summative assessment of teacher effectiveness at the expense of formative improvement.** Many schools in the United States have adopted accountability-driven reforms to their teacher evaluation systems, emphasizing standardized test data and teacher value-add models (Croft et al., 2015), and this was indeed the case in DC public schools (National Research Council, 2015), as well as many area charter schools. In DCPS, the important role of instructional coach may have been lacking in many schools, and in any case, standardized test data are not particularly trusted, useful, or timely for teacher self-improvement (Economic Policy Institute, 2010; Ingram et al., 2004). Although we attempted to help fellows build their evaluation plans on top of existing school procedures as advised (Donaldson et al., 2016), this effort may have pushed fellows (especially math) towards use of standardized data they already collected. On the other hand, additional data collection may burden teachers and students.
3. **Real-time and interim opportunities for data analysis, reporting, and reflection may help teachers apply findings to their teaching**. We initially assumed the summer was the best time for teachers to compile and reflect on results; however, due in part to year-to-year changes in fellows' schools and courses, and the limited availability of the past year's data, fellows did not seem to agree. Rather, they desired real-time, ongoing feedback to inform their instruction immediately. The inclusion of a midterm assessment and report (perhaps completed over the holiday break) may have helped meet this demand, even for fellows who focused on standardized test data. Engaging with program and school mentors and teams

may have helped promote ongoing reflection when evaluators were unavailable (Donaldson et al, 2016), in addition to encouraging fellows to share interim growth with their students.

4. **Producing aggregate, summative evaluations of program impact may still be unfeasible, given the wide variety of data collected**. As the discussion of results above suggests, despite our best efforts, fellows' individual results were not readily quantified and combined to provide an overall story of program impact. We could have enhanced aggregation by requiring particular outcomes and instruments (Saxton et al., 2014), although there is no consensus on favored methodologies (Isore, 2009). These challenges are common in evaluations of STEM teacher training programs (Alemdar et al., 2018; Hardre et al., 2014). We attempted to strike a delicate balance between top-down and bottom-up standards and needs for evaluation, but we tended to prioritize fellows' perspectives (see below).
5. **Evaluators need to meet teachers where they are to adapt approaches to their needs**. Each teacher is on a unique process of learning, as our results suggest. Some fellows had great comfort and familiarity with basic assessment and statistical procedures, in part through required coursework and professional development, but also their past experiences and inclinations. Others obviously struggled to meet our requirements. However, we found that our individualized approach to fellow self-evaluation: promoted buy-in and intrinsic motivation, despite the external mandate (Firestone, 2014); produced results that were personally meaningful and useful (Smagorinsky, 2014); and helped develop each fellow's evaluation capacity in specific areas of need (Archibald & Vo, 2018). In addition, the flexible approach permitted fellows to design mini-experiments to guide their particular practice, as discussed above, and to collect data on multiple outcomes of interest.

Conclusion

Our innovative process of teacher self-evaluation touched on many tensions inherent to teacher training evaluations and the broader evaluation field: tensions between accountability and learning, between evaluations mandated from the top down and developed from the bottom up, and between the use of standardized data to measure student content mastery and a host of other data sources and student outcomes worth investigating. Ultimately, we navigated these tensions by attempting to build the evaluation capacity and evaluative thinking of the participating teaching fellows (Archibald & Vo, 2018), in hopes that fellows would internalize these skills and become critically reflective teachers capable of ongoing self-reflection and self-improvement beyond the program (Archibald et al., 2018; Brookfield, 1995). Our preliminary results suggest our approach has the potential to be successful, at least in the short term.

Besides offering practical guidance and lessons learned to evaluators for building evaluation capacity in an accountability-driven context, this article contributes to the paucity of scholarly research on evaluation of STEM teacher training programs (Hardre et al., 2014; Saxton et al., 2014), especially of Noyce teacher scholarship programs (Alemdar et al., 2018). Future studies could be more intentional about identifying and tracking specific indicators of evaluation capacity and evaluative thinking over the long term (Archibald & Vo, 2018), possibly drawing from our inductive themes.

Further, we did not explore the complex question of how the approach might contribute to or hamper efforts to increase equitable outcomes for historically disadvantaged schools, teachers, and students (Donaldson et al., 2016; Carden, 2007), although the setting for this work was in a school district with a majority of disadvantaged students, and fellows' self-reflection touched on issues of equity (e.g., exploring different outcomes for ELL and SPED students). For now, however, we hope this article helps spark conversation within the broader evaluation field

about how to promote reflective practice through self-evaluation, and the difficulties of implementing and legitimating the practice in contexts dominated by accountability concerns.

References

Alemdar, M., Cappelli, C. J., Criswell, B. A., & Rushton, G. T. (2018). Evaluation of a Noyce program: Development of teacher leaders in STEM education. *Evaluation and Program Planning*, *71*, 1-11.

Archibald, T., Neubauer, L. C., & Brookfield, S. D. (2018). The critically reflective evaluator: Adult education's contributions to evaluation for social justice. In A. T. Vo & T. Archibald (Eds.), *Evaluative Thinking. New Directions for Evaluation, 158,* 109–123.

Archibald, T., & Vo, A.T. (Eds.). (2018). *Evaluative Thinking. New Directions for Evaluation, 158*.

Buckley, J., Archibald, T., Hargraves, M., & Trochim, W.M. (2015). Defining and teaching evaluative thinking: Insights from research on critical thinking. *American Journal of Evaluation, 36(3),* 375-388.

Brookfield, S. (1995). *Becoming a Critically Reflective Teacher.* San Francisco: Jossey-Bass.

Carden, F. (2017). Building evaluation capacity to address problems of equity. In S. Sridharan, K. Zhao, & A. Nakaima (Eds.), *Building Capacities to Evaluate Health Inequities: Some Lessons Learned from Evaluation Experiments in China, India and Chile. New Directions for Evaluation, 154*, 115–125.

Croft, S. J., Roberts, M. A., & Stenhouse, V. L. (2015). The perfect storm of education reform: High-stakes testing and teacher evaluation. *Social Justice*, 70-92.

Donaldson, M. L. (2012). Teachers' perspectives on evaluation reform. Washington, DC: Center for American Progress.

Donaldson, M. L., Woulfin, S., LeChasseur, K., & Cobb, C. D. (2016). The structure and substance of teachers' opportunities to learn about teacher evaluation reform: Promise or pitfall for equity? *Equity & Excellence in Education*, *49*(2), 183-201.

Dweck, C. (2007). *Mindset: The New Psychology of Success.* New York: Ballantine Books.

Economic Policy Institute. (2010). Problems with the use of student test scores to evaluate teachers. EPI Briefing Paper# 278. Washington, DC: Economic Policy Institute.

Firestone, W. A. (2014). Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher*, *43*(2), 100-107.

Guskey, T. R. (2018). Does pre-assessment work? Educators must understand the purpose, form, and content of pre-assessments to reap their potential benefits. *Educational Leadership*, *75*(5), 52-57.

Hardré, P. L., Ling, C., Shehab, R. L., Herron, J., Nanny, M. A., Nollert, M. U., ... & Wollega, E. D. (2014). Designing and evaluating a STEM teacher learning opportunity in the research university. *Evaluation and Program Planning, 43*, 73-82.

Henderson, C., Beach, A., & Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching, 48*(8), 952-984.

Ingram, D., Louis, K. S., & Schroeder, R. G. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record, 106*(6), 1258-1287.

Institute of Education Sciences. (2013). The effectiveness of secondary math teachers from Teach For America and the Teaching Fellows Programs. NCEE 2013-4015. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Isoré, M. (2009). Teacher evaluation: Current practices in OECD countries and a literature review. OECD Education Working Papers, No. 23. OECD Publishing.

McLaughlin, J. E., Roth, M. T., Glatt, D. M., Gharkholonarehe, N., Davidson, C. A., Griffin, L. M., ... & Mumper, R. J. (2014). The flipped classroom: a course redesign to foster learning and engagement in a health professions school. *Academic Medicine, 89*(2), 236-243.

National Research Council. (2015). *An Evaluation of the Public Schools of the District of Columbia: Reform in A Changing Landscape*. Washington, DC: The National Academies Press.

Noyce, P., Perda, D., & Traver, R. (2000). Creating data-driven schools. *Educational Leadership*, 57(5), 52-56.

Ozogul, G., & Sullivan, H. (2009). Student performance and attitudes under formative evaluation by teacher, self and peer evaluators. *Educational Technology Research and Development, 57*(3), 393-410.

Ross, J. A., & Bruce, C. D. (2007). Teacher self-assessment: A mechanism for facilitating professional growth. *Teaching and Teacher Education, 23*(2), 146-159.

Saxton, E., Burns, R., Holveck, S., Kelley, S., Prince, D., Rigelman, N., & Skinner, E. A. (2014). A common measurement system for K-12 STEM education: Adopting an educational evaluation methodology that elevates theoretical foundations and systems thinking. *Studies in Educational Evaluation, 40*, 18-35.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference, 2nd ed.* Boston: Houghton Mifflin.

Smagorinsky, P. (2014). Authentic teacher evaluation: A two-tiered proposal for formative and summative assessment. *English Education, 46*(2), 165-185.

Table 1: Guiding Questions for Fellow Evaluation Plans

Question #1: What is your plan for credibly measuring student outcomes in your class?
a. Do/can you implement pre/post assessments? If so, is this approach department-wide or your own?
b. Will you look at average or individual student improvement? By class or across classes?
c. What other quantitative indicators will you collect?
d. What qualitative data will you collect?
e. How will data be stored and analyzed?
Question #2: How will data help you better understand your teaching and student learning?
a. How will you utilize the data to help improve your performance as a teacher?
b. How and when will you reflect on evaluation results and adjust your teaching in response?
c. Will you apply data differently for different students/classes?
d. What if any external audiences will you share results with (e.g., administration, other teachers)?

Table 2: Original Guiding Questions for Fellow Student Outcomes Reports

Part 1: Student Outcomes Results
a. Share the results of your pre/post assessments of student content mastery.
b. Provide additional quantitative data that would be helpful for interpreting the results.
c. Summarize qualitative data that would be helpful for interpreting the assessment results.
Part 2: Teacher Reflection Results
a. Describe 3-5 examples (1-2 paragraphs each) of how you utilized the data to help improve or demonstrate your performance as a teacher
b. What changes will you make to your Evaluation Plan for the upcoming school year?
c. Any final comments or feedback on the student outcomes evaluation process, either related to the previous school year or looking ahead to next school year?

Figure 1: Theory of Change of STEM Teaching Training Program

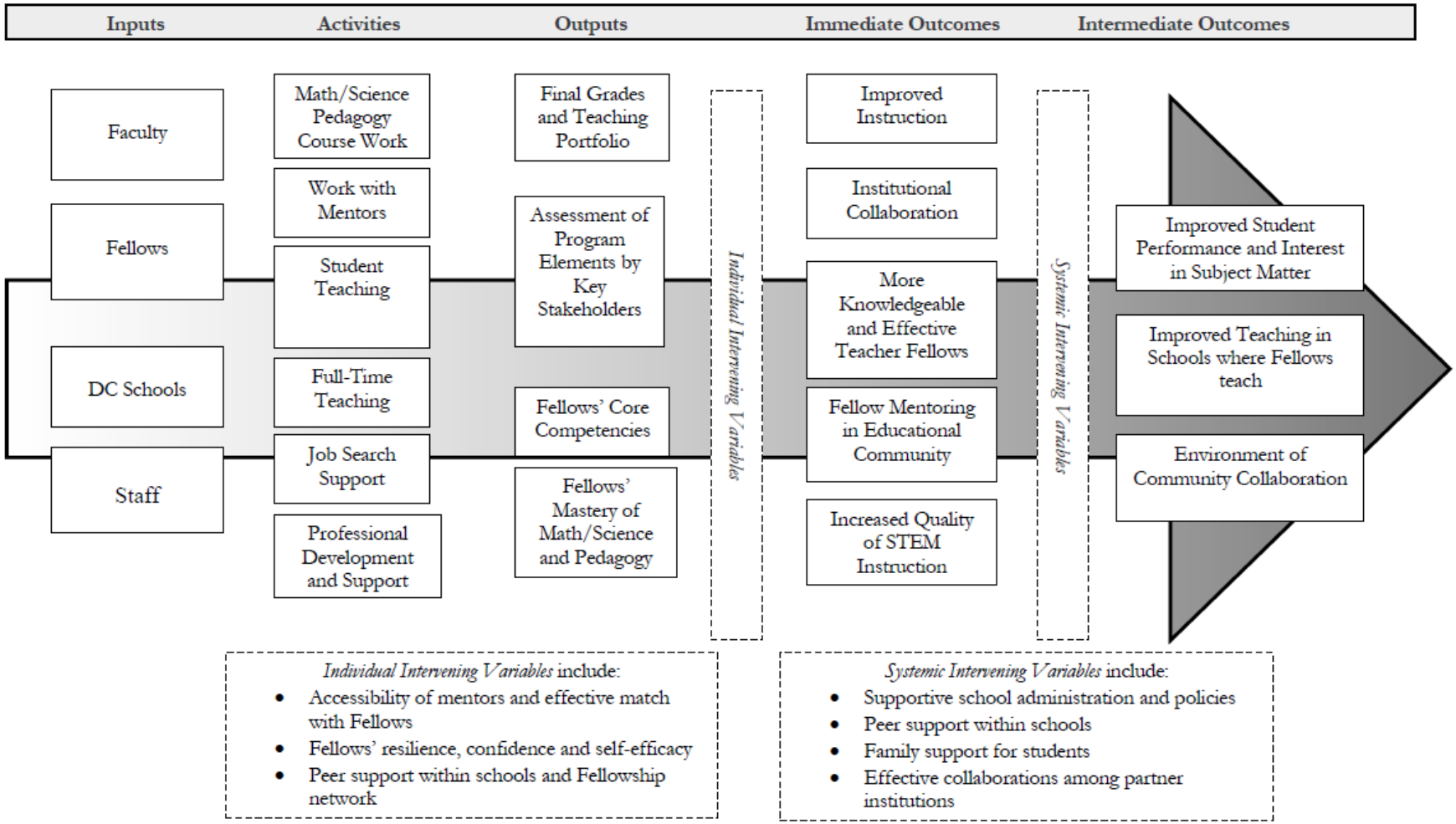
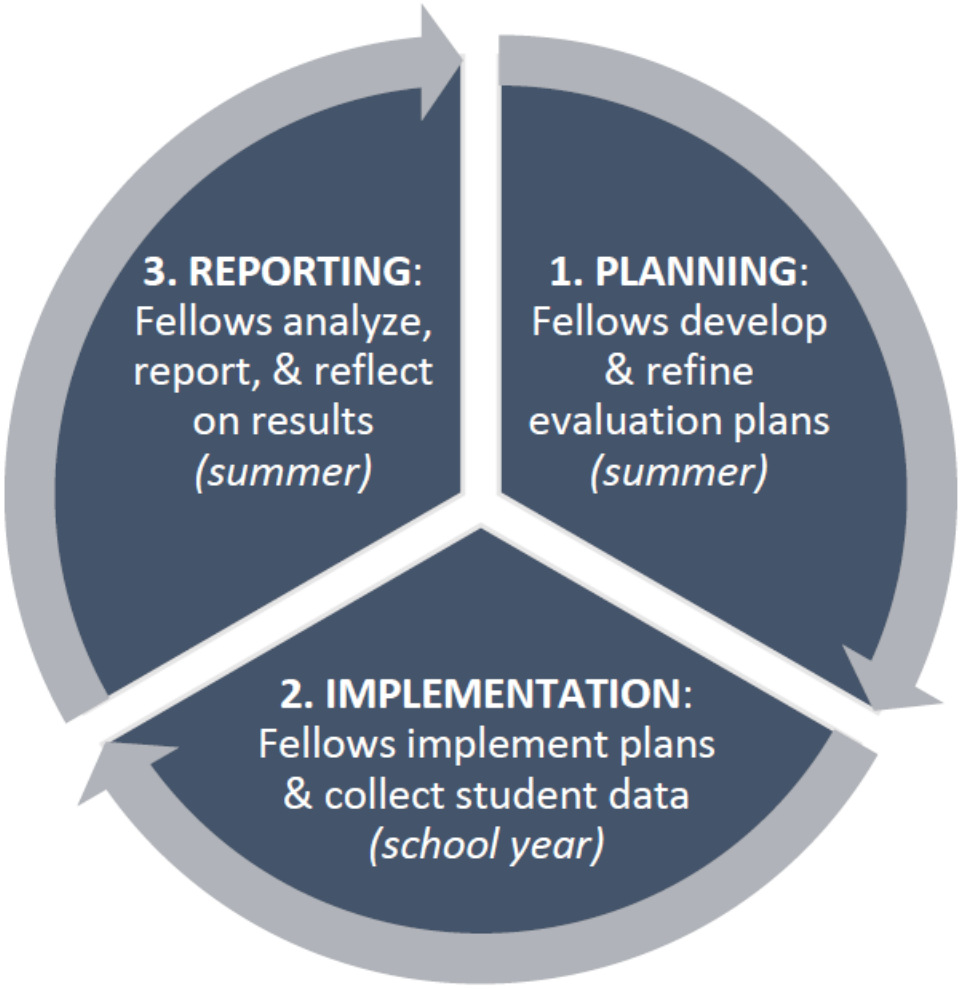


Figure 2: Cycle of Fellow Self-Evaluation



SUPPLEMENTARY FILE
Fellow Student Outcomes Evaluation:
Final Results Reporting Template

Name:

School Name: ☐ **DCPS** ☐ **Charter** ☐ **Other_____**

Grade(s): ☐ **High School** ☐ **Middle School**

Subject(s): ☐ **Science** ☐ **Math**

Please submit your results summary for the previous school year by **August 1** to… At that point, the evaluators will review your summary and either approve it or work with you to address any questions or concerns we have. If you need any guidance on preparing your results, please don't hesitate to contact… Be sure to refer back to your initial **Evaluation Plan**, and **proposed modifications** from last year's report, as needed. *Thanks for your participation!*

APPROX. TOTAL NUMBER OF STUDENTS TAUGHT THIS YEAR:

WHAT ASSESSMENT STRATEGY(IES) DID YOU EMPLOY (*check all that apply*):

☐ Self-created or adapted assessment(s) ☐ School/dept-wide comprehensive/final exam(s)

☐ Standardized exam(s) (e.g., AP, iReady, MAP) *Please provide name(s)*: _____

☐ Other *Please describe*: _____

ON WHAT TIMELINE DID YOU COLLECT ASSESSMENTS (*check all that apply*):

☐ Pre *Month(s) collected*:_____ ☐ Post *Month(s) collected*:_____

☐ Interim *Month(s) collected*:_____ ☐ By Unit *# of units & dates*:_____

☐ Other *Please describe*: _____

PART I. Student Outcomes Results

A. Please share the results of your pre/post assessments of student content mastery. You may want to provide a simple table, along with 1-2 paragraphs explaining your results. See the following page for a sample table that you can adapt as needed. ***Please provide the raw data files to us via email if possible, preferably as an Excel document.***

Group	**Assessment Used** *(title or brief overview)*	**PRE Assessment Average Score** (n=)	**POST Assessment Average Score** (n=)	**Average Change Score** *(statistical significance at p<.05 noted with *)*	**Score Range, Median**	**Notes** *(e.g., caveats in data collection/ analysis)*
Group A **(n=X)** *(include brief description: subject, age, level, etc.)*						
Group B **(n=X)** *(include brief description: subject, age, level, etc.)*						
Group C **(n=X)** *(include brief description: subject, age, level, etc.)*						
Group D **(n=X)** *(include brief description: subject, age, level, etc.)*						
Group E **(n=X)** *(include brief description: subject, age, level, etc.)*						

B. Please provide any additional quantitative data that would be helpful for interpreting the results above AND RELATE THESE DATA TO YOUR ASSESSMENT RESULTS ABOVE, to the extent possible (e.g., through correlation, sub-group analysis, or eyeballing patterns in data). These data may include final grades, attendance, student survey results, etc. You can organize these data in the same fashion as the pre/post assessment results (e.g., by class or subject), and feel free to use tables or graphs in addition to text. ***Please provide the raw data files to us via email if possible, preferably as an Excel document.***

WHICH TYPES OF QUANTITATIVE DATA ARE YOU REPORTING *(check all that apply):*

☐ Student grades or pass/fail rates ☐ Student survey responses (closed-ended)

☐ Attendance data ☐ Behavior data ☐ Other *Please describe*:_____

C. Please summarize any qualitative data that would be helpful for interpreting the assessment results. For example, you might paraphrase student open-ended feedback, describe illustrative student projects or other sample work, or provide stories illustrating student outcomes.

PART II. Teacher Reflection Results

A. WHICH IF ANY CHALLENGES DID YOU EXPERIENCE IN YOUR DATA COLLECTION?

☐ Missing student data (e.g., because of student turnover, attendance/behavior issues)

☐ Lack of student effort on exams (e.g., because exam did not count for grade or credit)

☐ Students permitted to retake exams to increase scores

☐ First year with new school, curriculum, and/or exam

☐ Unique student characteristics for all/some classes (e.g., ELL, SPED)

☐ Students entering school year behind/below grade level

☐ Other *Please describe*:_____

Please pick one or more of the challenges you indicated above, and provide more detail about the challenge(s). How did it affect your evaluation process and/or results? Any lessons learned or ideas for overcoming the challenge in future years?

B. HOW DID YOU USE YOUR EVALUATION RESULTS *(check all that apply):*

☐ Identify areas of student strength and weakness (e.g., for review/tutoring, differentiation)

☐ Modify lessons/curriculum (e.g., adjusting presentation style, level)

☐ Modify assessments (e.g., item analysis)

☐ Shared results/growth with students (e.g., to motivate them, identify improvement areas)

☐ Shared results with collaborating teacher(s) (e.g., for planning lessons and interventions)

☐ Shared results with dept head/administrators (e.g., as part of teacher evaluation)

☐ Other *Please describe*:_____

Please pick one or more of the uses you indicated above, and provide more detail. When and how did you use the evaluation data and results in the specified way? Was it used for formative improvement during the school year, and/or summative assessment and future planning after the school year? Did you experience any challenges or barriers to the specified use, and if so, any lessons learned or ideas about how to overcome them?

C. WHAT ARE YOUR PLANS FOR SELF-EVALUATION, CONTINUOUS LEARNING, AND REFLECTIVE PRACTICE GOING FORWARD? In other words, how will you continue to assess your teaching and students' learning, and improve your practice, beyond the fellowship program and minimal school requirements? Do you intend to continue applying this self-evaluation process on your own, or do something totally different, and why? This is also your opportunity to provide any feedback on the benefits or shortcomings you experienced with this evaluation process, and how you might improve on it.