Supporting rigorous evaluations: An evaluator’s perspective

Julie Edmunds, SERVE Center at University of North Carolina at Greensboro

A paper presented at the 2016 Fall Research Meeting of the Association for Public Policy Analysis and Management, Washington, DC

It was at the first i3 project director’s meeting in 2012 when it well and truly hit me that the Investing in Innovation (i3) program represented an entirely new set of federal expectations relative to evidence and evaluation. The entire focus of that meeting was on the requirements for the evaluation, with multiple sessions on designing an impact evaluation that met What Works Clearinghouse standards and on measuring Fidelity of Implementation. The project directors attended most of the same sessions and heard the same things I heard. At one point during the meeting, they turned to me with an expression of disbelief and said, “You really do this all day?” I smiled and said, “Yes, isn’t it fun?” They did not necessarily agree that it was fun, but they did leave the meeting with an increased appreciation of the importance of the evaluation in this project.

This paper presents my experiences as an evaluator with multiple i3 projects. A little different than a traditional research article, this paper is more of a narrative of a personal evaluation journey. I begin by relating the experiences and relationships that led to the submission of an i3 proposal, move to the description of the services provided to support the i3 evaluation and then conclude with a discussion of the personal and organizational impacts of those services.

The path to an i3 proposal

My involvement in the i3 world had its roots in a long-established relationship with a project developer and in efforts to study a specific model they were implementing—the early college model. In 2004, I was hired at the SERVE Center at the University of North Carolina at Greensboro with the intention of half of my time being spent providing research support for the New Schools Project (New Schools), a new public-private partnership that was attempting to re-imagine the high school experience in North Carolina initially through small, theme-based schools. One of the first things I was asked to do was to summarize the research on how school reforms fail (Edmunds, 2005). As Tony Habit, the organization’s executive director said, “I want to know how to avoid the mistakes that others have made.” I also summarized research on topics such as the small schools work and served overall as a critical friend; the New Schools’ staff did not necessarily always like the information I gave them (especially if it did not support the direction in which they wanted to move) but they did appreciate it and often modified their approach based on the information I provided.

One of the high school reform models being implemented by New Schools was early college, small schools that blurred the line between high school and college (Jobs for the Future, 2008). I watched as the first schools were being established in North Carolina as small schools of choice—similar to magnet schools—and noticed that many schools had more applicants than they had spots, an ideal situation for a randomized controlled trial. At that point, almost two
years into our relationship, I had established some trust with New Schools and they agreed to support a research study to determine the effectiveness of the early colleges. Starting in 2006, I led a multi-institutional research team (including researchers from Abt Associates, Duke University, and later RTI International) funded by the Institute of Education Sciences to conduct a prospective longitudinal experimental study of the impact of the early college. Participating schools agreed to use a lottery to select their students and the study followed results for students who got in through the lottery and those who did not.

Although we experienced some initial challenges in recruiting schools to use a lottery (documented in Edmunds, in press), we ended up with a robust sample of 4,000 students who applied to 19 early colleges. As the study progressed, our team provided regular updates to New Schools, sharing the impact results and also providing more formative data to help influence the design. For example, we found that early college students struggled in some settings in mathematics, at least partially because the schools were so small that math teachers had to teach multiple topics in math instead of specializing in one or two topics as often happens in larger settings. As a result, New Schools developed professional development and coaching targeted to math teachers.

Overall, the study found positive impacts for early colleges on almost all of the outcomes we examined, including success in a college preparatory course of study, attendance, suspensions, and remaining in school (Edmunds et al., April, 2011; Edmunds et al., 2012; Edmunds, Willse, Arshavsky, & Dallas, 2013). Early college students also reported more positive school experiences, including more support, more rigorous and relevant instruction and better relationships with their teachers (Edmunds, Willse, Arshavsky, & Dallas, 2012; Edmunds, et al., 2013). Given the strong research design and the positive results, we received two additional consecutive IES grants to continue to follow the students. In those studies, we found that early college students enrolled in postsecondary education and received postsecondary credentials at a higher rate (Edmunds et al., in press).

With these positive findings, New Schools was interested in expanding the model beyond small schools of choice. They sought to apply the principles of the early college to comprehensive high schools and applied for an i3 Validation grant to test whether the model worked in these settings. The research from our longitudinal experimental study done on their early college work was cited as core support for grant. SERVE was included as the evaluation partner and the project was funded as part of the 2011 competition.

After this first grant was awarded, other organizations, including Jobs for the Future (JFF) who had a long history with early college work, also applied for funding seeking to implement the early college model in multiple settings. Our experimental research as well as a subsequent experimental study by AIR that utilized some of our data (Berger, Turk-Bicakci, Garet, Knudson, & Hoshen, 2014; Berger et al., 2013) was included as justification for the grants. JFF was awarded funding under the 2012 competition for a project being implemented in Denver and in the Rio Grande Valley of Texas. Teachers College at Columbia, with the support of JFF, then received a grant for STEM-oriented early colleges in Michigan and Bridgeport, Connecticut in 2013. In the 2014 competition, New Schools was awarded a Scale-up grant to implement early college efforts in five states and finally, Columbus State Community College, partnering with...
JFF, continued the perfect streak of funding early college-related projects with an award in 2015. I kept thinking that there would early college fatigue in the i3 world and they would stop awarding grants but there has been continued interest. As one of the few evaluators working on early colleges and because of my established relationships with New Schools and JFF (I had previously done a research project with them on why students leave early colleges), I served as the lead evaluator on all five of those projects. My exposure to the i3 world thus began with an award made in the second round of competition in the winter of 2011 and will continue through 2020 when the 2015 grant concludes.

In addition to the original early college study, my relationship with New Schools had also included a small descriptive evaluation that looked at the implementation of their principal professional development efforts. However, neither this evaluation, nor even the longitudinal experimental study, adequately prepared all of us for the i3 evaluation requirements.

i3 evaluation expectations and supports

The 2011 i3 Request for Proposal included language that indicated that the proposal would be assessed based on the quality of the evaluation or the “The extent to which the methods of evaluation will include a well-designed experimental study or a well-designed quasi-experimental study” as well the extent to which the evaluation would examine implementation and provide feedback and information to allow for later replication of the project (U.S. Department of Education, June 3, 2011). The RFP also indicated an expectation to participate with technical assistance that would be provided to ensure a high quality evaluation. I had previously seen similar language with another federal grant for which I was an evaluator in the Teaching American History project, which had an expectation of an experimental or quasi-experimental design and which provided a small amount of technical assistance at the project directors’ meeting. However, that particular grant program had nothing like the follow up and teeth that accompanied i3.

The i3 program was authorized in its legislation to include the sort of accountability for evaluation missing in other federal programs. Federal grant programs are required to report on a set of Government Performance Results Act (GPRA) indicators. The i3 evaluation represents the first time (of which I am aware) that two of these program GPRA indicators related to the quality of the evaluation. Usually these indicators related to the impact of the program such as people served or improved outcomes of some sort; indeed, the i3 program has two of these types of indicators, the number of students served and the cost per student. However, there are also two indicators with language specifically related to the evaluation:

(2) the percentage of programs, practices, or strategies supported by a Validation grant with ongoing well-designed and independent evaluations that will provide evidence of their effectiveness at improving student outcomes; (3) the percentage of programs, practices, or strategies supported by a Validation grant with ongoing evaluations that are providing high-quality implementation data and performance feedback that allow for periodic assessment of progress toward achieving intended outcomes (U.S. Department of Education, June 3, 2011, p. 32170)
These indicators meant that the program and, by extension, the grantees were being held accountable for the quality of the evaluation.

Abt Associates was the contractor hired to help develop and implement a process for getting the grantees and evaluators to meet these indicators. They did this with three interrelated activities: 1.) establishing a format and structure for the evaluation plans and their review; 2.) establishing a set of standards to guide review of these plans; and 3.) providing ongoing technical assistance to help in meeting those standards. Below I discuss these activities and my interactions with them.

Format and structure of the evaluation plans. The first formal requirement for the evaluation was to submit an updated plan to the project officer within 90 days of the award being made. This evaluation plan was intended to expand on what was submitted in the proposal but, in the earlier grants, there was no guidance given as to what should be included in this submission. Evaluators were then expected to submit a much more detailed evaluation plan by October 1st in the first year of the project—approximately 9 months after the grants were officially awarded. These detailed evaluation plans were reviewed by Abt’s Analytic and Reporting (A&R) team to determine whether or not they met the i3 evaluation standards (discussed below). After the review process, the A&R team provided detailed feedback as well as an overall assessment of whether the design met the evaluation standards.

One of the first tools that Abt provided to the evaluators was a Design Summary template delineating the content of the detailed evaluation plan. The Design Summary asked the evaluators to pre-specify all of the components of the impact plan and a subset of the components of the implementation plan. Evaluators were not required to use the Design Summary structure but it was strongly encouraged because the Design Summary included all of the components that were going to be subject to review. There were sections in the Design Summary that asked the evaluators to describe the intervention, specify the research questions, define the outcomes, specify the samples for each outcome and the time in which they will be collected, and describe the analytic strategy.

Accompanying the Design Summary was a spreadsheet tool entitled the Contrast Tool. This tool assisted the evaluator in clearly specifying the specific contrasts that would be included in the impact evaluation. Completing the Design Summary and Contrast Tool forced the evaluation team (often with the project team) to be very clear around when outcomes needed to be assessed and which students would be included in which sample.

The Design Summary also included space for the evaluator to provide a formal logic model, identify the Key Components of the intervention and describe how fidelity of implementation would be assessed.

The final tool that the evaluation team was asked to complete was a Fidelity of Implementation (FOI) matrix. This matrix was a template in which evaluators were asked to identify the Key Components of the project and the level of implementation that was required to be considered as successful implementation. Completing this template required regular and frequent interactions with the developer, pushing them to clearly define the activities that they were going to do and
the level at which they were going to be implemented. For example, all five of our i3 projects had a component that focused on coaching of participants. For the FOI matrix, we had to work with the project staff to clearly define what counted as a coaching activity and to determine what would be an acceptable level of coaching. Some readers might expect that these would be the types of things that would be specified prior to the grant being awarded; however, we found that in all five of the projects, project staff had to have lengthy conversations with each other and with us to allow those levels to be set.

The evaluation standards. To guide the review of the evaluation plans, the Abt team established evaluation standards for both the impact evaluation and the implementation evaluation.

One of the goals of i3 is to expand the number of high quality impact evaluations that pass the What Works Clearinghouse (WWC) review process. As a result, Abt developed a set of explicit impact evaluation standards for the i3 program that were very closely aligned to the WWC standards (Institute of Education Sciences, n.d.). These evaluation standards were shared with the evaluators in a monograph, through webinars, and in professional development sessions and were used to guide the review process.

All of the evaluations for our five i3 projects were similar in design. They used a quasi-experimental design in which schools that were implementing early college strategies were matched to comparable schools not implementing the intervention. As a result, we knew that the highest WWC rating we would be able to attain would be WWC standards with reservations.

Over the course of the i3 program, the WWC standards were revised (Institute of Education Sciences, March, 2014) with some clarifications that affected baseline equivalence for cluster-level studies such as ours. This affected the review process and resulted in our second evaluation plan being identified as not meeting the i3 standards, although, in the year prior, the exact same approach had been identified as meeting the standards. The specific issue had to do with whether or not a previous cohort of students could be used to establish baseline equivalence if the analytic approach was a multi-level model. Although it was frustrating that the standards had changed over time, we were able to respond to the review process and justify some of our decisions and change some of our other plans to accommodate the review; as a result, the revised plan met the standards the second time. We also revisited the first evaluation plan and modified it to accommodate the revised standards.

Although the Abt team had been trained as WWC reviewers, they did not actually work for the WWC and this led to some challenges with interpretation as the WWC review guidance does not take into account all possible research scenarios. This was particularly salient for our cluster-level quasi-experimental designs that incorporated outcomes besides standardized tests. As a result, the Abt team often was trying to predict how WWC reviewers might respond to specific issues and we had to occasionally negotiate with them about what might be an appropriate interpretation.

For example, the WWC guidance and i3 guidance both indicated that outcomes had to be defined and collected in the same way between the treatment and control groups, a necessary precursor for a high quality impact study. For our study on early college work in North Carolina
comprehensive schools, we wanted to use graduation rates as an outcome—a theoretically and practically important outcome. WWC guidance suggested, however, that graduation rates could only be used as an outcome if there were identical graduation requirements in each district or if the schools were all located in the same district with the same requirements. Because our study was a whole school reform effort being implemented in rural districts, both of these stipulations were impossible to meet. Graduation standards did vary across districts in both our treatment and our control groups. Additionally, the districts were often small and contained only one high school or had all of their high schools participating in the project so comparison schools would not have been able to be identified within the district. As a result, the Abt team postulated that we would not meet WWC standards relative to this outcome. This became a true source of frustration to me because I believed that this interpretation of the WWC standards would mean that the only quasi-experimental whole school reform studies with a graduation rate outcome that could ever meet standards would be ones that looked at high schools in urban districts with lots of schools. We pushed back with the reviewers, arguing that consistent definition is not required if the definition of the outcome was exogenous to the intervention and if there was no systematic difference in how the outcome was defined between the treatment and control groups. We even documented the graduation requirements for the treatment districts and our likely comparison districts to indicate that there was no systematic pattern in definitions between the two groups. The i3 review team agreed with our rationale; we have not yet received a WWC review but believe that it will pass the outcome standards.

The WWC does not have standards relative to implementation evaluations so the Abt team had to develop their own standards. These standards were fairly straightforward and included the creation of a detailed logic model and the identification of a process to assess the fidelity of implementation of the program’s Key Components. We faced no trouble meeting the implementation evaluation standards but have heard that enough other projects faced challenges that the Abt team added an “interim” review where the evaluators could make modifications in response to the formal Abt review before a final assessment was made.

Technical assistance. Abt also provided extensive technical assistance in helping to interpret the WWC guidance and assisting the evaluators in meeting the standards. They provided professional development sessions that accompanied each of the project directors’ meetings. Additionally, they provided a series of webinars on the i3 standards as well on particular methodological topics with which multiple grantees appeared to be struggling, such as matching techniques. These sessions and webinars were useful as they were usually based on real-life scenarios faced by grantees.

The primary technical assistance resource, however, was the TA liaison for the project. These were very experienced methodologists who were assigned to each project and who reviewed drafts of the Design Summary, Contrast Tool and FOI matrix and provided feedback. In my case, I had the same TA liaison for all five of the evaluations. I was very fortunate in my allocation because my liaison was also the lead analyst on our longitudinal experimental early college study. Therefore, we had a pre-existing, well-established relationship and he already had deep knowledge of the intervention. Our TA liaison therefore became essentially another member of the research team who helped us think through some thorny methodological issues. He was extremely thoughtful and would be honest enough to identify where the i3 standards might be a
minimum standard that may not be high enough to reach best practice in research. He would also identify times where he did not have the answer and would reach out to other members of the Abt team to find us guidance.

At the beginning of the projects, we would have regularly scheduled meetings to ensure that we were on track for completing the Design Summary. However, we could always contact him at any point if we had specific questions or just wanted him as a sounding board. Our involvement with the TA liaison was never threatening; instead it felt as if someone else was invested in the success of our project. My experience with the TA liaison was certainly colored by the fact that I had worked with him on other projects but I have spoken to many other evaluators who also commented on the high quality support that they received from their liaisons.

Impact of the i3 Evaluation Process

Overall, participating in the i3 evaluation technical assistance has had a significant impact on me personally, on my organization, and, I believe, on the projects themselves.

Personally, the i3 work taken as an entire package has been some of the best professional development I have experienced. Through the tools and TA support, our research team has developed an in-depth understanding of the WWC standards. I have also greatly expanded my own expertise in quasi-experimental design as we sought to understand and apply the WWC standards to our particular circumstances. The many evaluation requirements for i3 could easily have felt like so many hoops to jump through; however, what the process mostly felt like was an effort to ensure that our evaluation was the most rigorous it could be. Even when we disagreed with the guidance coming from the Abt personnel, it always felt like a productive discussion among research professionals where disagreement was going to lead to a stronger product.

The i3 evaluation work has also benefited my broader university community. I have done a university-wide workshop on the evolving federal expectations regarding program evaluation and am planning a graduate-level course that will center on designing education evaluations to meet the WWC standards.

The i3 evaluation has also changed how I work with clients in several ways. First, the i3 project requires an independent evaluation and clearly delineates what it means for an evaluation to be considered independent. Evaluation is generally a somewhat odd beast because it is most commonly paid for by the project but is expected to provide unbiased information about the project. The well-defined expectations around independence made it easier to have some of the tougher conversations with the clients. Moving forward with non-i3 projects, I will likely incorporate similar language into my agreements with my clients.

Second, the i3 process definitely raised the awareness of project staff of the importance of the evaluation. The project staff developed an increased understanding of the merits of experimental and quasi-experimental designs and became more educated consumers of evaluation. This made my job as the external evaluator much easier.
Finally, the evaluation tools created for i3 also improved the implementation of the projects. I had always worked with the developers to create a fleshed out logic model of the project and had found those conversations extremely helpful in terms of crystallizing the developer’s intent. However, I had not previously worked with them to take the next step and create such a detailed plan for assessing fidelity of implementation.

The FOI matrix has been the part of the evaluation process that was probably most directly helpful to the project staff. Completing the FOI matrix required that the evaluator and client work together to clearly specify the project activities and the appropriate threshold level for those activities (how much is enough?). This made the project staff more purposeful in their planning. One of our clients used the FOI matrix as a project planning tool because it made them be very detailed and transparent about the activities that were expected to be completed as part of the project. They found it especially useful in facilitating clarity and agreement among the different partners working on the project.

As the project was being implemented, the FOI matrix also served as an accountability tool for the project staff. Even though many of the measures in our projects were at a relatively simplistic level that is often thought of as bean-counting (i.e., # of days of professional development days), tracking these measures helped the project staff ensure that their activities were being completed as planned.

Supporting the i3 evaluation effort has required a significant commitment of resources on behalf of the federal government. At the end of the i3 experiment, there will be many evaluators who have increased their skills. There will also be many project developers who have improved their understanding of evaluation. There will also be the uncounted ancillary benefits from interactions these evaluators and project staff have with others. From an evaluator’s perspective, these impacts show that the resources have been well spent.

References


