**Decision-Focused Evaluation: Testing a Framework for Education Evaluations**

*Fellow: Dominic Zarecki, Fortune School of Education*

***Synopsis****: As the lone data wolf at Fortune School of Education, a network of charter schools in California, SDP Fellow Dominic Zarecki was tasked with evaluating an intervention for over 500 students in grades K through 8 during 2017-18. He used this experience to develop and test a framework for planning and executing evaluations. This framework helped Zarecki determine how and when to make a radical yet reasonable recommendation: switch to a new intervention in 2018-19.*

**The Challenge**

In 2017 Margaret Fortune, CEO of Fortune Schools, decided to prioritize helping all students improve their reading fluency: the ability to read both quickly and accurately. Some leaders on her executive team had argued that a lack of reading fluency is a key educational barrier; students who struggle to understand a text may be less willing or able to engage in more complex learning activities. Therefore, Fortune hired teaching aides to implement a “fluency program” to support all students below the national median of fluency for their grade level. Principals and aides at the six Fortune schools had wide latitude to determine the details of implementation.

Zarecki was given the role of evaluator for this fluency program. While he wanted to accurately measure its impact, organization leaders were more interested in obtaining evidence-based recommendations to improve the program over time. Leaders had prioritized reading fluency, and they were ready to make evidence-based changes in pursuit of that goal. How could Zarecki use the tools of evaluation to make recommendations for particular adjustments? This challenge led him to pilot an “intervention” in his own work: a framework called Decision Focused Evaluation (DFE).

Program evaluation is the process of systematically using information to estimate the effectiveness and efficiency of interventions or policies. Evaluation is a necessary but insufficient condition for data-informed decision making. To be useful, evaluations must fit within the constraints faced by decision makers. Even a perfectly executed randomized controlled trial has limited usefulness if it occurs after a budget decision is made, or if it is unclear how the results of an intervention compare to viable alternatives. How can data strategists ensure their evaluations are useful?

**The Intervention**

Hasbrouck and Tindal (2006) report national norms about the number of words per minute the median student is able to read at specific points in each grade of each school year. If Fortune Schools were perfectly average, half the students would be above and below the median student. At the beginning each trimester, aides used grade-level texts to measure the number of words students could read accurately in one minute. Since some Fortune schools focused on reading fluency in previous years, it is not too surprising that 60% of students were above grade level on fluency at the beginning of the 2016-17 school year. (To be clear, this doesn’t reflect general academic achievement; Fortune students currently score slightly below average on nationally normed ELA assessments.)

*Details about Implementation*

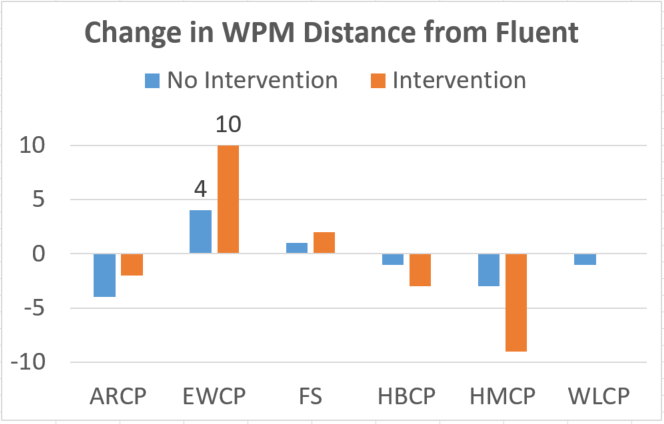
Zarecki used informal data gathering such as emails and conversations to learn that the fluency program was implemented in very different ways across sites and grades. Typically, students were pulled from their ELA class for individual or small group instruction. This occurred anywhere from two to five times a week for 15 to 60 minutes per meeting. Students in grades four and higher usually worked through a program called Rewards. The Rewards curriculum is designed to be a 6-week course – taking 50-60 minutes of in-person instruction each day – where students learn strategies that help improve reading fluency, comprehension, and academic vocabulary. However, time constraints meant that students did not always receive the entire course. Students in grades three and under did not have any prescribed materials, so instructors had discretion to customize lessons, often based on factors such as student interest. Instructors vary in their background and experience, but typically are salaried aides who love teaching and currently lack formal teaching credentials.

*First Decision Point: Identifying Outliers to Improve Implementation*

By the first week of December, students took another fluency assessment to measure how many words per minute they now could read accurately. There was not time to conduct a regression discontinuity analysis. Instead, Zarecki generated quick descriptive statistics. He looked at students near the cutoff for receiving the intervention – within 10 words per minute. The 104 students who barely made the intervention did not improve any more than the 140 students who barely missed it. He also looked at students at one site where staff did not use the test score to determine who received the intervention. Zarecki found that students who were supposed to receive the intervention improved more when they did not receive the intervention than vice-versa. While not ideal, it was enough to conclude that our intervention was not improving fluency. (A subsequent regression discontinuity analysis confirmed this.)

Unfortunately, stopping the intervention altogether was not an option. Fluency was an organizational priority, posted all over the walls of the school buildings. Instead, Zarecki recommended trying to improve the intervention. He utilized a Tableau dashboard to look for patterns and found that one school, Ephraim Williams College Prep, made progress in reading fluency – especially with students who received the intervention (see Figure 1).

***Figure 1: Descriptive analysis uncovers outlier school***



*Notes: WPM = words per minute. Each acronym represents a different school.*

This school also took a very different approach to the fluency program. Instead of using aides to pull students out of class, they made the ELA teachers implement the fluency program with their entire lower-ability classes. This different implementation could have caused the school to be an outlier, so Zarecki recommended that all schools conduct the fluency program as a full class intervention. Starting in second grade, Fortune schools sort students into ELA classes by ability, so aides and teachers would work with groups of students in the lower ability classes. The CEO embraced the idea, so that change occurred during Trimester 2 after winter break.

*Second Decision Point: Responding to Impacts*

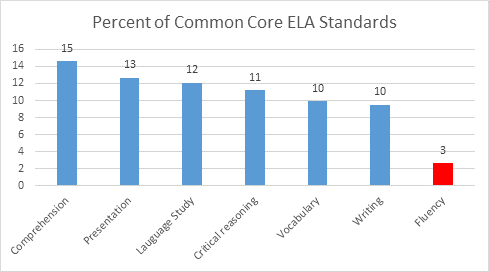
Unfortunately, this adjustment did not appear to improve the fluency program. By every metric, reading fluency decreased over trimester two. The percent of students at or above median fluency dropped steadily: from 60% at the beginning of the year, to 58% at the end of trimester 1, to 55% at the end of trimester 2. During trimester 2 students only improved an average of 9 words per minute – less than the 12 words per minute they gained last trimester. Even if a regression discontinuity analysis found positive effects, it is problematic that overall fluency metrics declined.

There is another, more fundamental problem with the fluency program: fluency improvement appears unrelated to academic growth. Comparing fluency improvement to growth on our NWEA ELA assessments reveals essentially no relationship; the correlation is 0.04. The correlation between fluency improvement and Smarter Balanced growth is negative 0.11. These national assessments do not explicitly test fluency. In fact, these assessments even try to avoid indirectly testing fluency by being untimed and providing an assortment of accommodations, such as text-to-speech, that allow students to engage the test without being fluent readers.

With additional adjustments, the fluency program may have improved overall fluency. Perhaps the dosage was too small, and students simply needed more time and/or individualized attention. Perhaps staff needed more rigorous materials and/or training on how to improve fluency. However, even a substantively effective fluency program would not change the fact that there is no relationship between fluency growth and ELA growth on national assessments. If the underlying goal is to support overall ELA improvement, then a focus on reading fluency will not move students in that direction.

It is difficult to recommend simply stopping an intervention. Instead, Zarecki recommended switching to an intervention that focuses on reading comprehension: the most prominent set of Common Core standards (see Figure 2). He used the What Works Clearinghouse to search for interventions that had proven effective in improving reading comprehension. Only one intervention had results that were both positive and statistically significant: Read 180.[[1]](#footnote-1) The CEO and executive team are currently considering his recommendation.

*Figure 2: Common Core Standards focus on Comprehension, Not Fluency*



**Impact**

Zarecki provided realistic and evidence-informed recommendations at two key decision points during the school year:

*Adjusted from a small-group to a full-class intervention*

Recommended that aides and teachers would work together to conduct the fluency program as a full-class intervention. This was based on an analysis by site which found that the fluency program was only successful at the one site that implemented the fluency program this way. The CEO embraced the idea, so that adjustment occurred after winter break.

*Considering a switch to an intervention targeting comprehension*

Recommended that the organization switch the priority from fluency to comprehension. This was based on the fact that (1) the fluency program did not improve overall fluency and (2) improvement in fluency was unrelated to growth on the ELA Smarter Balanced summative assessment. This idea is still under consideration by the CEO and executive team.

Perhaps more importantly, Zarecki created a framework, available in the appendix, that he and other data strategists can utilize when planning and executing evaluations.

**Lessons Learned**

1. Decision-makers may need information sooner than you think. Zarecki initially believed he would be able to conduct a regression discontinuity analysis during winter break and communicate the results in early January. However, this timeline was thrown off when the CEO needed to make a decision at an executive team meeting in mid-December. It is worth thinking through contingency plans: what evidence can you utilize if you need to make a recommendation before you are able to conduct your full analysis?
2. Think creatively and early about how to evaluate realistic intervention adjustments. Zarecki wished he had talked to aides during the summer before the fluency program began. If he had asked them how they planned to implement it, he would have realized how much variation there would be. If he could re-do things, Zarecki would survey all aides about the details of their implementation. He also would have tried to find some aides willing to do small-scale experiments by trying two different implementation strategies.

**Practical Advice**

Prioritize understanding the decision-making context. There may be important deadlines that are not well known, such as draft budget estimates or contract deadlines. Additionally, unexpected events such as new research or staff changes may alter which options seem reasonable to decision-makers.

To get focused project time, try to “swap” work and vacation days. Working even a few days while everyone else was on winter break provided an amazing opportunity to devote uninterrupted time to the evaluation.

Collect more information about the intervention than you think necessary. It is much easier to gather information while the intervention is happening than to do so retrospectively. Additionally, try to gather information in a structured way (e.g. surveys with 5-point scales) so that you can measure implementation attributes in a comparable somewhat rigorous manner.

Look in detail at the studies in the What Works Clearinghouse. Zarecki found that even when filtering specifically for reading comprehension, many interventions only had impacts in related domains. Additionally, the vast majority of studies were small RCTs that found positive but insignificant effects. While this is promising, it is only limited evidence in favor of those interventions.

**Questions to Consider**

Think about an intervention or program in your own organization. The first step of Decision-Focused Evaluation is to answer questions about the audience, timing, and options of each decision-making context:

1. Foundations: What data will you have, and how will the evaluation be conducted?
   1. What data is relevant to evaluating the intervention? How frequently will you collect that data, and how accurate will it be? Does the team or person conducting the evaluation know how to plan and conduct it? Can that team or person influence the implementation of the intervention to ensure that it is compatible with a rigorous evaluation? What processes will ensure that the evaluation is accurate and as rigorous as possible?
2. Audience: Who needs to approve changes to the intervention?
   1. Even if one person is formally in control, who will that person trust when considering a change? Do staff on the ground have authority to adjust the intervention or even make more substantial changes within their limited context? Are people willing to change the intervention if the evaluation finds that it is not obtaining the intended results?
3. Timing: When will your audience make decisions about changing the intervention?
   1. Do contracts, budgets, meetings, or other events impose deadlines on these decisions? When there is flexibility, changes should occur based on the schedule of data collection – assessments, surveys, etc. If you backwards plan from those dates, when will decision-makers need to see your evidence?
4. Options: What realistic changes could your audience make to the intervention at that time?
   1. It’s almost always possible to make tiny adjustments to an intervention. What specific adjustments are reasonable given your constraints and theory of action? Alternatively, can you scale the intervention? Switch to a new intervention? Always ask yourself: what would be the best way to achieve our goals?

**Appendices**

1. Decision-Focused Evaluation:
   1. Infographic
   2. Toolkit
   3. Template

**Decision-Focused Evaluation (DFE) Infographic**

**STEP 1: The decision-making context**

The relevant evidence depends primarily on the options for change:

*\*RCT = Randomized controlled trial; RDD = Regression discontinuity design; PSM = Propensity score matching*

**STEP 3: Execution and communication**



**Protect Obtain Communicate**

**Time Details Clearly**

**Decision-Focused Evaluation Toolkit**

Program evaluation is the process of systematically using information to estimate the effectiveness and efficiency of interventions or policies. Evaluation is a necessary but insufficient condition for data-informed decision making. To be useful, evaluations must fit within the constraints faced by decision makers. Even a perfectly executed randomized controlled trial has limited usefulness if it occurs after a budget decision is made, or if it is unclear how the results of an intervention compare to viable alternatives. How can data strategists ensure their evaluations are useful?

**The Framework: Decision-Focused Evaluation**

Decision-Focused Evaluation (DFE) is a framework to help data strategists maximize the impact of their work. The goal is to structure our efforts in order to provide rigorous and relevant insights at crucial decision points. For any intervention or policy, the framework has three steps:

1. Outline the decision-making context for a particular intervention
2. Determine the most rigorous evaluations possible within that context
3. Execute and communicate those evaluations as convincingly as possible

The first step of DFE is to determine the decision-making context for a particular intervention. This can be summarized as answering three questions:

1. Audience: Who needs to approve changes to the intervention?
   * Even when one person formally “owns” an intervention, usually multiple people need to be in agreement to approve a change. For example, a superintendent may choose to only make changes when the chief academic and financial officers are on board with it. When considering intervention adjustments (rather than switches or changes in scale), power may be explicitly devolved to a number of staff on the ground. In the extreme case, each interventionist may need to be convinced individually to make particular adjustments.
2. Timing: When will decisions about changing the intervention need to be made?
   * This is almost never referring to the formal point at which boards approve budgets or plans are finalized. While that may be when decisions become official, people make the actual decisions well before then. Precision is not vital; if all you know is that the relevant discussions occur in May, then you should plan to communicate your results by the end of April. When there is flexibility concerning when decisions can be made, it is best to time changes to maximize the ability to measure impact. For example, school systems that administer interim assessments at the end of each trimester should make changes that take effect at the beginning of a trimester.
3. Options: What realistic changes could be made to the intervention?
   * In broad terms, there are always three types of changes we can make to an intervention. First, we can shift the scale of the intervention. This type of change usually is a viable option at particular points of the year; it just requires that an intervention can be expanded or removed from certain sites or grades. Second, we can switch to a new intervention. This typically is only viable in the long term because of the operational difficultly of launching a new intervention; it takes time to obtain materials, train staff, etc. Finally, we can adjust the intervention itself. Virtually always a viable option, this might be as minor as tweaking the implementation process or as major as hiring and training staff in a new role specifically designed to support the intervention.

The second step is to use the future realistic options to determine what data and analyses are needed to make the best possible recommendations:

1. Shift the scale of the intervention
   * This first type of change involves two closely related decisions: in which direction should we shift the scale of an intervention, and by how much? To make this decision, we want to know the average effect of the intervention; ideally, we would even know the average effect for particular sites, grades, and subgroups. The direction and size of the effect (positive or negative) should be directly related to the direction and size of the shift in scale. We should estimate the average effect size with the most rigorous possible analysis: ideally a randomized controlled trial, second best is regression discontinuity, and third best is a regression-based analysis such as propensity score matching.
     1. Randomized controlled trials require the creation of treatment and control groups. How are you going to randomly administer the intervention to only some of the people who could receive it? When this is possible, it’s only necessary to have the relevant outcome data about treatment and control groups. The actual analysis is a simple comparison of treatment and control groups. However, running a multivariate regression and including a set of independent variables can improve the precision of estimates and make it more likely the results are statistically significant. (Note: School staff tend to be most comfortable when both treatment and control groups receive some form of intervention. When this occurs, the evaluation compares those interventions to each other. If these interventions are focused on distinct goals – such as English vs. Math – then it is essentially equivalent to conducting two experiments simultaneously.)
     2. Regression discontinuity analyses requires using a cutoff to at least help determine who receives an intervention. Is there a cutoff that people cannot manipulate that could be used to at least recommend that people receive an intervention? To be rigorous, you will need to collect a variety of independent variables to test whether people just above and below this cutoff are similar. This analysis is statistically complicated, but there are packages that make it relatively simple to execute the core analysis, such as *rdplot* for Stata. (Note: School staff tend to be very comfortable using cutoffs!)
     3. Propensity score matching is necessary when the above analyses are not possible. It requires either a control group or variation in the treatment. While you do not need to randomly create the control group or variation in the treatment, you do need to try to include independent variables that reflect the types of factors that might support alternative hypotheses about the findings. For example, if students who received more treatment also were more likely to have disabilities, then it is possible that disability status is causing the treated group to have relatively low academic growth. You would need to control for disability status in your analysis to remove that alterative hypothesis as a possibility. (Note: Propensity score matching has the unfortunate problem of being both the least persuasive and most confusing of all these analyses.)
2. Switch to a new intervention
   * This second type of change involves the decision to stop the current intervention and start an entirely new intervention. Even when switching to a new intervention is operationally possible, doing so in an informed manner requires estimating the impact of at least one intervention that you have not attempted. This requires reading someone else’s evaluation, gauging its accuracy, and determining whether a similar effect is likely to occur in your context. To determine the relative costs, it will also be necessary to obtain information about cost for both your own intervention and the possible new intervention.
3. Adjust the intervention itself
   * This third type of change almost always is a viable option. However, it is often difficult to find rigorous evidence in favor of a particular adjustment. At a minimum this requires measuring variation in the implementation of the intervention, probably through surveys, interviews, focus groups, or detailed observations. Even if this variability is measured, it is often difficult to prove that such variability caused the intervention to be more or less effective. The best approach is to support any quantitative findings with qualitative evidence from the field as well as corroborating evidence from the existing literature.

The third step is to plan and execute the most rigorous evaluation that can be done by the deadline. While there are many important pieces to this process, the ones below are both particularly important and not things that are taught in typical graduate programs.

1. Protect project time
   * For education data strategists, there is never enough time to do everything we would like. When deadlines are pressing, you inevitably will have to prioritize your workload and ignore some tasks on your to-do list. Three strategies help to prevent evaluations from getting dropped. First, secure internal allies – especially your supervisor and their boss(es) – who will push for the successful completion of the evaluation. A second, related strategy is to find external stakeholders – such as a grant funder or powerful interest group – who will pressure your organization to produce the evaluation as a deliverable. Lastly, reserve as much calendar time as possible to working on the evaluation. You are lucky if other urgent work only consumes half of the time officially “reserved” for a particular project!
2. Obtain detailed evidence
   * Data strategists often have training that focused on mastery of quantitative analyses. While these are vital to your work, the systematic collection of detailed quantitative and qualitative evidence is at least as important. Imagine yourself defending your recommendation at a decision point. How could a skeptic poke holes in your argument? Try to collect the evidence you would need to counter those arguments. First, make every effort possible to obtain baseline data. Second, try to administer surveys and/or conduct structured observations of the intervention. Third, supplement your qualitative and quantitative evidence with relevant existing research when necessary.
   * Note: It is always good to seek expert feedback on your evaluation; such feedback is especially important for portions of your evaluation about which you are most concerned or uncertain.
3. Communicate clear results
   * Graduate school rarely trains people to speak simply. While complex, technical details are important to uncovering truth (and should be dealt with and documented!), those details need not always be communicated when sharing that truth. What is the highest level thrust of your finding? Can you tell a short story about a particular example (e.g. a real or hypothetical school, class, or student) to explain what you found?
   * When necessary, clearly explain the information a rigorous evaluation can provide – and what positive or negative results would imply about the impact of an intervention. Also, manage your own expectations. Leaders face a variety of pressures and do not always embrace the recommendations of evaluations, even ones conducted by top-tier experts. Make all ethical efforts to push your recommendations, but be prepared for your organization to not implement a significant portion of them.

**Decision-Focused Evaluation Template**

*Instructions: Fill in all blue text with information from your own project.*

***Step 1: The decision-making context***

|  |  |
| --- | --- |
| **Audience**  *Who approves and influences changes to the intervention?* | *e.g. Superintendent in consultation with Executive Team; Principals have discretion over certain aspects of implementation.* |
| **Timing**  *When will people decide about any changes?* | *e.g. Agenda packets for a crucial meeting are finalized by Jan 15; adjustments would ideally be made soon after testing on Oct 31 and/or Mar 1.* |
| **Options**  *What realistic changes could they make at that time?* | *e.g. Can adjust mid-year; can shift to more or fewer grade levels in year two; and consider piloting a new intervention no earlier than year three.* |

***Step 2: Data collection and analyses***

|  |  |  |  |
| --- | --- | --- | --- |
| **Realistic Option** | **Shift the scale of the intervention** | **Switch to a new intervention** | **Adjust the intervention itself** |
| **Learning Goal** | Average treatment effect, ideally also for specific subsets: schools, grades, subgroups, etc. | The likely impact of a new intervention, in your context | The impact of variation in the implementation of the intervention |
| **Data and Method** | Randomized controlled trial if you can randomly assign to treatment and control; Regression discontinuity design if you can use a cutoff to at least partially determine treatment; and if the above options are unavailable, use propensity score matching with as many relevant control variables as possible. | Read evaluations about other interventions, gauge the strength of their impact on the outcomes you care about, and determine whether a similar impact is likely given the specifics of your context. (What Works Clearinghouse is a great starting point.) | Measure as many details about implementation as possible through surveys, interviews/focus groups, structured observations, and new quantitative data collection. Try to conduct micro-experiments to test the impact of the most salient implementation variations. |

***Step 3: Execution and communication***

|  |  |
| --- | --- |
| **Protect project time** | Find internal allies and external stakeholders who will push the evaluation as a deliverable; reserve lots of calendar time for working on the evaluation. |
| **Obtain detailed evidence** | Think like a skeptic! Obtain baseline data, use surveys or structured observations, and supplement with existing research. |
| **Communicate clear results** | What is the high-level conclusion? Can you explain it with a story? |

1. For the Read 180 page on What Works Clearinghouse: <https://ies.ed.gov/ncee/wwc/Intervention/742>. For the Read 180 product page: <https://www.hmhco.com/products/read-180/>. [↑](#footnote-ref-1)