GOLDILOCKS DEEP DIVE

Introduction to Rapid-Fire Operational Testing for Social Programs

FEBRUARY 2016

Acknowledgments: This Goldilocks Toolkit was authored by Mary Kay Gugerty, Dean Karlan, Tetyana Zelenska, with editing and design by the IPA Communications Team (David Batcheck, Laura Burke, Jennifer Cowman, Heidi McAnnally-Linz, Megan McGuire).

ipa
INNOVATIONS FOR POVERTY ACTION

Introduction to Rapid-Fire Operational Testing for Social Programs

One key message of the Goldilocks Initiative is that impact evaluation is not for everyone. Yet, even when measuring impact is not feasible, social enterprises and non-profits can still answer important questions about their programs using rigorous measurement techniques.

One of these is techniques is rapid-fire testing:[1] randomized trials that compare the effect of related interventions on a single, immediate (or short-term) outcome. This method is used to test operational issues and aims to influence immediate outcomes, such as product take-up, program enrollment, loan repayment, and attendance, among others. In rapid-fire tests, participants are randomized into different treatment groups (and sometimes, but not necessarily, a pure control group) and exposed to variations in a program's design or message.

The outcome of interest (usually program take-up or use) is measured and compared across treatment and control groups. Often outcomes are measured administratively, so that there is no large survey undertaking necessary in order to gather the data. For example, tests may use data from a financial institution that the institution would gather anyhow (deposits, loan repayments), from a store on sales, or from online tests tools such as Google Analytics and Optimizely, which facilitate both the test and the data collection.[2]

Before discussing the when, why, and how of rapid-fire testing, we run through an example from Barack Obama's 2008 presidential campaign. The campaign used its website extensively to sign up potential supporters, who it tapped for volunteers and contributions.
The campaign team used rapid-fire testing to increase the number of people who signed up on the website.

While the method can be useful for answering questions about the response to a design tweak, in most cases, it does not measure the impact of a particular intervention—whether the intervention made people better off compared to how they would have been without it. Before considering if rapid-fire testing is an appropriate tool for a specific program, it is important to understand when to use the method, its advantages and limitations, and some basic requirements for successful testing. It is also helpful to learn from the experiences of other programs that have used it. Below we offer

[1] What we refer to as "rapid-fire" testing is sometimes referred to by others as A/B testing, particularly in the marketing field, where it originated. Technically, "A/B testing" implies testing across two groups: A and B. It may be a short- or long-term study; and A and B may be different variations of the same intervention, or A may be a treatment and B may be a control. However, A/B testing is commonly used in marketing and implies both rapid-fire and subtle or small variations within a larger intervention or policy or product. When there are more than two variants, it is often called "bucket" or "split" testing.
[2] The following website is helpful in deciding which A/B testing software to use: http://www.conversion-rate-experts.com/split-testing-software/

guidelines on these points, provide a checklist of the steps involved in implementing rapid-fire tests, and share examples from two organizations that have applied the method to their work.

1. Question: how can the campaign website maximize the number of people who sign up? Prior to the test, the website had an 8.6 percent sign-up rate.

2. What to test: the team tested many ways to increase sign-ups:

- Four sign-up buttons
  "Sign up" *"Sign up now"* *"Join us now"* *"Learn more"*
- Six different visual media (photos and videos)

Combining each button with each image/video resulted in 24 different home pages to test (4 buttons x 6 media)

3. Testing: Every visitor to the campaign website was randomly directed to one of the 24 versions of the home page. The team tracked the views and sign-ups for each combination.

4. Results: The "Learn More" button + a photo of the Obama family performed best. It had a sign-up success rate of 11.6 percent; 40 percent more than the original rate.

The team estimated the sign-ups from this tweak resulted in an additional:

- 2.8 million email addresses
- 288,000 volunteers
- $60 million in contributions

Rapid-fire testing delivered fast, reliable results on how to optimize website messaging, allowing the campaign team to act quickly to increase sign-ups, and ultimately, volunteers and contributions.

From Siroker, Dan. "How Obama raised $60 million by running a simple experiment." *The Optimizely Blog*. November 29, 2010. https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

When to Use Rapid-Fire Testing

Rapid-fire testing is most suited for answering questions that generate a fast feedback loop and for which administrative data are recorded. Because it answers questions about program design, it is particularly valuable in the design or pilot stage, or when expanding a program to new areas or new populations. It can provide credible insights into program design, produce highly actionable data, and do so at relatively low cost.

Rapid-fire testing is a useful complement to traditional monitoring and evaluation activities: monitoring tracks a program as it is implemented and can provide information on how to improve program performance, while impact evaluation assesses a program as it was implemented. Rapid-fire testing can be used to modify a program's *design*, where monitoring improves *implementation*. The method provides rigorous evidence on how the design of a program affects take-up and use, and eliminates the need to rely on guesswork or trial and error.

Rapid-fire testing can be especially useful for answering questions about the early stages of a program's theory of change. Theories of change rely on a number of explicit and implicit assumptions about how a program will work. Early-stage assumptions describe the links between activities and outputs, such as demand levels for a product or service (will people enroll or buy a product?). Whether or not these assumptions hold often depends on how information is conveyed to or received by the target population. Rapid-fire testing can be used to investigate these assumptions to see which design features, marketing or messages increase the take-up and use of a

new program or product. See the Executive Summary on building a theory of change for more discussion on theories of change and assumptions.

For example, programs often face the challenge of assessing demand for a product or program before they launch or scale to new areas. Rapid-fire testing can be used to identify what small and often costless design or messaging change increase demand, and thus, product sales or program participation. Answering questions like these early in program implementation enables the design of a better product prior to a full-scale roll-out, and may help prevent waste of organization's resources in areas where demand is low. Fast feedback is key. For rapid-fire testing to work as intended, outcomes must occur immediately after the program tweak is made or message delivered, so that differences can be more confidently attributed to the change.

Requirements for Rapid-Fire Testing

Certain basic requirements must be met for rapid-fire testing to be feasible and useful:

A Program or Service That Can Be Varied: Rapid-fire testing compares outcomes for different versions of a product or message. But not all products or services can be easily varied. Sometimes changes are too difficult or too costly to implement for a rapid-fire test, or the changes they seek to cause are too slow to appear. For instance, a sanitation campaign that promotes hygiene practices via community engagement and education sessions may not be a good candidate for rapid-fire testing—altering the content of an education campaign is difficult and changing hygiene practices takes some time. On the other hand, a program that shares price information with farmers via SMS may be a good candidate for rapid-fire testing. Altering the messages that deliver the information (such as personalizing them, or framing them in either a positive or negative way) is easy to do, and should result in immediate action (farmers sell where prices are highest on a given day).

The Right Question: Questions amenable to rapid-fire testing have the ability to generate immediate action, such as "how does the phrasing of my invitation email affect who signs up for my online service?" Because of this, rapid-fire tests are often very context specific and have low transportability to other settings. Rapid-fire testing is typically not suited for answering questions about welfare changes, mainly because welfare changes take time and usually is not captured by administrative data. Additionally, a program's impact on wellbeing needs to be measured against a control group that does not participate in the program. Because most rapid-fire tests use current clients or people already engaged in an intervention, a pure control group is often not available.

Sample Size. As with any evaluation, rapid-fire testing requires sufficient sample size to confidently determine the effect of each treatment. Variations in the product design or messages being tested will likely result in small and incremental effects. This means that a large sample size may be needed to achieve sufficient statistical power for an organization to have confidence in the results. Depending on the platform and scale, this is less of a concern for online campaigns and services, which can reach large numbers of individuals at low marginal cost. A number of tools can be used to

calculate appropriate sample sizes, such as Stata and Optimal Design—but these require a fair amount of statistical capability. Sample-size constraints may limit the viability of rapid-fire testing during the pilot of a new campaign or intervention, when the program is rolled out to a limited number of participants.

Data Systems. While implementing a rapid-fire test is relatively straightforward, accurate results depend on credible data and require systems that can manage the rate and volume of incoming data. Rapid-fire testing has the potential to produce very large datasets—particularly if data are high-frequency or if there are a large number of observations, such as the number of page views, phone calls, or SMS from thousands of cell phone subscribers. To ensure that these data are credible and can be acted on in a timely manner, an organization must have adequate systems to receive and process the data, checking for completeness, consistency and quality. See the [Deep Dive on Using Administrative Data for Monitoring and Evaluation] for a discussion of these challenges and for possible solutions.

Analytical Capacity. Rapid-fire testing requires basic analytical skills to perform the power calculations that determine the minimum required sample size and statistical training in hypothesis testing. Online resources can assist with both of these tasks[3]. Additionally, most of the A/B testing software for websites will automatically calculate the results once key indicators are entered (such as the number of overall webpage visitors and the number of overall conversions, for A and B options) and will report whether it is statistically significant or not.[4]

Steps in Conducting Rapid-Fire Tests

Step 1: Define the question of interest: This operational question guides the design of the rapid-fire test. For example, an organization seeking to increase donations may ask, "how can we change our messages to make them more compelling?"

Step 2: Identify the sample and randomization strategy: As with any evaluation, it is necessary to identify the population to include in the test and determine how to randomly select individuals into treatments. For example, to test the impact of different email messages on donor contributions, an organization will need a database of eligible users or participants. Randomizing based on IP address is a common approach for testing messages on a website. If randomization is not possible, then rapid-fire testing is not a viable way to answer the question.

Step 3: Clarify which elements of the product or service can be controlled: An organization must be able to control the elements it intends to test. Identifying these elements often requires input from the implementation team, and involves assessing the feasibility of making the change and the actionability of resulting data. If a program cannot commit to implement one of the changes

[3] See Resources and Tools for Impact Evaluation (http://www.poverty-action.org/publication/resources-and-tools-impact-evaluation) for guides to conducting power calculations.
[4] For example, https://vwo.com/ab-split-test-significance-calculator/

under consideration, then it should not test that change. Doing so diminishes the statistical power of the test and wastes resources that could be better spent on something actionable.

Step 4. Prioritize what to test: The next step is to prioritize the potential changes by considering the theory behind them and by weighing the costs of implementing them against their expected benefits. For example, it is helpful to look at evidence-based theory when considering tweaks in messaging. Anchoring refers to the common tendency to rely on the first piece of information offered (the "anchor") when making decisions;[5] many marketers take advantage of this tendency when suggesting donation amounts. Similarly, an organization may be considering small design tweaks to its donations page (such as changing the color of the donate button to make it more visible) against the possibility of creating an entirely new donations page. Weighing the costs of each option against the expected benefit (informed by theory, where available) will allow an organization to make a responsible choice in the design of its rapid-fire test.

Step 5. Conduct power calculations and finalize treatment options: The next step is to perform power calculations to make sure the sample size has enough participants to measure the targeted outcome. Programs like Optimal Design can help determine the sample size. And web sites such as www.whichtestwon.com can also be a useful resource or reference: it includes a library of different website and email tests and their effects. Keep in mind that, holding the sample size constant, as the number of options being tested increases, statistical power is reduced.

Step 6. Run the test: Once the treatment options are finalized, it is time to randomize different treatment options to the sample of participants. Typically, this involves randomly splitting a list of participants into groups and assigning different treatments to each. The process of doing randomization itself is fairly straightforward – it can be done in Excel, statistical programs like Stata and SPSS, and websites like randomizer.org.

Step 7. Track the response: Organizations using rapid-fire testing will need an adequate data management system to track the outcomes from the test. A data management system links the user lists with the treatment assignment(s) and monitors outcomes. The quality of the results depends both on whether the assignment protocol is correctly followed and on the quality of incoming data. See [Deep Dive on Admin Data] for more information on administrative data.

Step 8. Analyze the data and apply the final results: The analytical capacity needed for this depends on the complexity of the research design. For simple web- or mobile-based A/B testing, the randomization and analysis may be automated through Optimizely or Google Analytics Content Experiments (GACE). These services tally all web visitors and then use randomization to split web traffic between the two versions of the webpage and provide descriptive statistics to compare which

[5] Kahneman, D. (2011). Thinking, fast and slow. Macmillan.

changes are leading to more signups. More advanced designs or tests that are not web-based will require greater statistical capacity.

Step 9. Consider replication: It is a good idea to replicate rapid-fire tests, both immediately and again over time. Consider a text messaging experiment that tested ten different versions of a text. Because the likelihood of false positives increases with the number of tests, replicating the test immediately can increase the robustness of the findings. And testing the messages again in a year can confirm that they are still effective, since it is possible that messages that worked well at first will lose their efficacy over time.

Rapid-Fire Testing in Practice

The cases below illustrate some of the ways rapid-fire testing can be used to answer operational questions. In both cases, the organization met the basic requirements for rapid-fire testing: the programs they sought to evaluate had components that could be varied, the questions they asked yielded quick answers, and they were able to attain sufficient sample size. And in each case, the organizations made sure they had sufficient analytical capacity by hiring an expert or partnering with an organization that specializes in impact evaluation.

Example 1. Improving Take-Up and Usage for a Business Mentorship Program

MicroMentor is a free social network that allows entrepreneurs and volunteer business mentors to connect with each other to solve problems and improve business growth. The organization has made over 5,000 matches between entrepreneurs and mentors since 2009 and expects another 1,000 matches in 2015. While this is an impressive number of matches, MicroMentor has struggled to measure their impact on business outcomes due to the challenge of constructing a valid comparison group.[6]

Even though impact evaluation may not be possible, MicroMentor has looked for rigorous ways to improve operations. MicroMentor operates exclusively through its online platform and wanted to optimize its website.[7] The organization hired a digital marketing expert to lead rapid-fire tests on key operational issues: take-up and engagement.

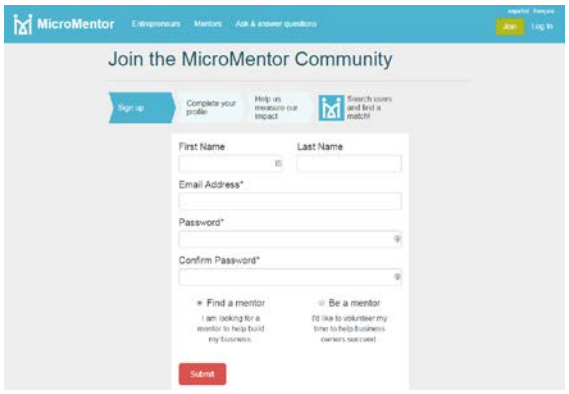
- **Take-Up**. Many MicroMentor site visitors who began the account registration process did not complete it, leading program staff to wonder how to encourage people to stick with the process. One possibility was that new users abandoned the registration process because it took longer than they expected. To test this, MicroMentor used Optimizely to conduct an A/B test that randomly directed some new users to a login screen with progress indicators (panel

[6] In the past, MicroMentor has conducted impact evaluations where they compared business outcomes among those who were mentored (found a match) and those who were not (but visited the website). However, issues of selection bias and potential low quality of self-reported data have led MicroMentor to abandon this method.

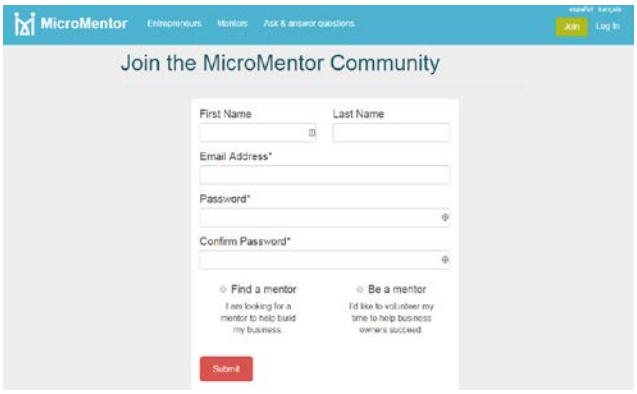
[7] Optimizing web usage refers to various procedures that make a website more popular in search engine results, such as adding relevant keywords and phrases on the website, adding tags and image tags, and optimizing other components of a website.

A), and directed other users to the original version with out the progress indicators (panel B). The team learned that giving more information on progress increased the likelihood of completing the process by twenty percent.

Panel A. New version of registration page with progress indicator



Panel B. Original version of registration page



- **Engagement**. MicroMentor believes that the quality and motivation of the program's mentors is an important element of program success. MicroMentor was interested in learning how to increase the mentors' motivation, which it measured by the number of mentors signing up for matches. The organization conducted a rapid-fire test that randomly selected some new mentors to receive a personal call from a member of the MicroMentor team. The rest of the mentors did not receive a personal greeting. Although the sample size for this experiment was relatively small (about sixty observations), the group that received a personal call resulted in a significantly higher number of matches. As a result, MicroMentor is considering introducing personal greetings of new mentors as part of its standard operational procedures.

Example 2. SMS Reminders to Save

Innovations for Poverty Action partnered with private banks in Bolivia and the Philippines to test whether simple SMS reminders could encourage people to save more.[8] The test involved sending different types of reminders to clients who had recently opened commitment savings accounts that had explicit savings goals. The test consisted of several treatment groups that received messages framing the savings goals in different ways (loss/gain framework) as well as a control group that did not receive a reminder. Table 1 shows the different messages in the treatment group for the sample in the Philippines.

[8] Karlan, D., McConnell, M. Mullainathan, S., & Zinman, J. (2016). Getting to the Top of Mind: How Reminders Increase Saving. *Management Science*.

Table 1. Test Summary for the Philippines

Timing	Goal mentioned?	Frame	Sample Assigned	Full Message
Regular only	"your dream"	Gain	163	Frequent deposits into the Gihandom Savings account will make your dream come true. A reminder from 1st Valley Bank.
		Loss	187	If you don't frequently deposit into your Gihandom Savings account your dream will not come true. A reminder from 1st Valley Bank.
Late and regular	"your savings goal", "your dream"	Gain	397	You didn't deposit in the 1st Valley Gihandom account for 30 days. Don't forget to deposit, so you can reach your savings goal, make your dream come true!
		Loss	410	You didn't deposit in the 1st Valley Gihandom account for 30 days. If you forget to deposit, you cannot reach your savings goal and make your dream come true!

Overall, the results indicated that the reminders were effective in helping clients to meet their savings goals. Messages that mentioned both savings goals and financial incentives (such as free life insurance in Bolivia) were particularly effective, while other content (such as gain versus loss framing and receiving additional late reminders) did not change savings behavior in any meaningful way.

The banks enacted some changes based on these findings. For example, the bank in the Philippines decided to continue sending SMS reminders because they proved to be very cost-effective, since the marginal cost of sending an additional SMS was nearly zero.

Conclusion

Operational testing with rapid-fire tests can be very helpful to organizations with questions on interventions that have a quick feedback loop, such as how users will react to different types of messages. Platforms exist that make the process of testing a limited number of variations fairly straightforward.

At the same time, the Goldilocks principles apply here: just because you *can* collect data does not mean that it is always the right use of resources. Rapid-fire tests are excellent options for certain measurement goals, but organizations must take care to use them credibly and commit to action based on the results. Rapid-fire tests are *not* designed to measure the final impact of an

intervention, and organizations must take care not to let the promise of quick results distract from impact evaluation priorities. Using administrative data may be inexpensive, but organizations must still ensure that it is credible. And, as with any other type of data collection, organizations should only test what they can commit to enacting. Data from rapid-fire testing should be used to inform decision-making.

For rapid-fire testing to be a responsible use of resources, organizations must plan carefully to make sure that the resources they devote and the changes they test are worth the costs. And they should be prepared for data that inform incremental steps, not game changers. Organizations that use rapid-fire testing may find it extremely helpful for testing the early stages of their theories of change and working out tweaks in program design before deciding whether overall impact evaluation is something to pursue next.

Appendix

Table A1. Difference between RCT and Rapid-fire Testing

	Traditional RCT	Rapid-Fire Testing
Question asked	What is the effect of encouraging savings on household welfare (psychological well-being, consumption, investments)?	What is the effect of sending a monthly SMS reminder on bank savings? What is the effect of a reminder that says "Please do not forget to save $X every month" versus a reminder that says "Please do not forget to save every month"?
Outcome of interest	School attendance, consumption, assets, investments, etc.	Savings in individual savings account
Treatment groups	A=randomly assigned to be encouraged to save B=randomly assigned to not be encouraged to save	A=randomly assigned to receive SMS reminder to save B=randomly assigned to not be encouraged OR A1=randomly assigned to receive message 1 A2=randomly assigned to receive message 2
Data source	Multiple; usually survey data and/or administrative data	Administrative data
Timeframe	Two years	Immediate to a few months