JANUARY 2014

Rigorous evaluation of financial capability strategies: Why, when and how

Perspectives from the field

cfpb Consumer Financial Protection Bureau

Table of contents

1. **Introduction: The need for rigorous evaluation of financial capability strategies 3**

2. **When is a randomized controlled trial (RCT) suitable? 6**

3. **How to conduct a strong study: Solutions to common challenges 9**

   3.1 Research design 9

   3.2 Research implementation 12

   3.3 Data collection and analysis 14

4. **Implications for practice, policy, and funding 17**

1. Introduction: The need for rigorous evaluation of financial capability strategies

The recent economic downturn raised awareness about the complexity of both our financial marketplace and the decisions consumers must make to manage their finances effectively. Despite the availability of a wide range of information about managing money and about financial products and services, many consumers still struggle to make the financial decisions that serve their life goals, a situation which can have significant long-term consequences for individuals and families. Therefore, helping consumers make well informed financial decisions that will serve them today and in the future is critical to the long term financial well-being of Americans.

The Dodd-Frank Wall Street Reform and Consumer Protection Act mandates that the Consumer Financial Protection Bureau (CFPB or Bureau) work to improve the financial literacy of American consumers. The Bureau is developing and implementing initiatives to educate and empower consumers to make better-informed financial decisions. This requires that we know what approaches are effective in improving financial decision making and financial well-being.

While the base of evidence regarding approaches aimed at improving financial decision making and outcomes (i.e. financial capability strategies) is growing, there remains too little rigorous empirical support. The implication of this is that service providers, financial institutions, policy makers, and funders have not been able to draw solid conclusions about which strategies are most effective. According to a 2011 Government Accountability Office (GAO) report on financial literacy, "[r]elatively few evidence-based evaluations of financial literacy programs have been

conducted, limiting what is known about which specific methods and strategies are most effective."[1] The Financial Literacy and Education Commission (FLEC),[2] comprised of 22 federal agencies and of which the Bureau's Director serves as Vice-Chair, also puts effectiveness at the top of its research agenda.[3]

In order to support and guide efforts to improve the effectiveness and quality of financial education, the CFPB is taking up this challenge to provide stronger evidence of what works to improve financial decision making and financial well-being, while ensuring the appropriate research protections for consumers. This effort will help us and a range of providers improve consumer decision making and outcomes. The Office of Financial Education, in coordination with the Office of Research, has developed a research program that focuses on (1) determining how to measure financial well-being, and identifying the knowledge, skills, and habits associated with financially capable consumers, (2) evaluating the effectiveness of existing approaches to improving financial decision making and outcomes, and (3) developing and evaluating new approaches.

As part of our research program to evaluate the effectiveness of existing approaches, the CFPB has contracted with the Urban Institute to engage in rigorous quantitative evaluation of promising financial education strategies, and to convene a peer learning network of other financial capability researchers and practitioners engaged in rigorous program evaluation using randomized controlled trials (RCTs) to measure program impact.

Given the promise that RCTs hold to produce the highest standard of quantitative evidence about the effectiveness of an intervention, but also given the difficulty of successfully implementing these studies in practical settings, the Urban Institute and the CFPB convened a roundtable discussion with a peer learning network on the benefits, challenges and best

---

[1] U.S. Government Accountability Office, GAO-11-614, *Financial Literacy: A Federal Certification Process for Providers Would Pose Challenges* (June 28, 2011), *available at* http://www.gao.gov/assets/330/320203.pdf.

[2] Congress established FLEC in 2003 with the mandate to improve the financial literacy and education of Americans, and to coordinate financial education efforts in the federal government. It is chaired by the Secretary of the Treasury.

[3] Financial Literacy & Education Commission, Research and Evaluation Working Group, 2012 Research Priorities and Research Questions, *available at* http://www.treasury.gov/resource-center/financial-education/Documents/2012%20Research%20Priorities%20-%20May%2012.pdf.

practices of conducting RCTs in the financial capability field in April 2013. In addition to Urban Institute and CFPB staff, 26 evaluators, funders, and program staff—all of whom were involved in ongoing or recent evaluations of financial capability programs—discussed their experiences and shared practical and useful insights into successful strategies and pitfalls of conducting rigorous evaluation of financial capability interventions.

The purpose of this report[4] is to share those insights with other researchers, practitioners, and funders undertaking or contemplating rigorous research into the effectiveness of different financial capability approaches. This sharing is critical because without a growing body of rigorous evidence of what works, assessments about whether the programs being offered will actually give consumers the skills and tools they need to make better financial decisions will continue to be inconclusive.

[4] This publication is based on a report to the CFPB prepared by Urban Institute researchers Brett Theodos, Margaret Simms, Claudia Sharygin, Rachel Brash, and Dina Emam under contract number CFP-12-Z-00006.

2. When is a randomized controlled trial (RCT) suitable?

RCT evaluations select units (individuals, schools, neighborhoods, etc.) at random from the same population, and assign them to one of at least two groups: treatment (sometimes called experimental) or control. Such a process helps to make the treatment and control groups equivalent—for example, with respect to motivation, ability, knowledge, socioeconomic and demographic characteristics, etc.—at the start of the study. Then, if all goes well, any differences in outcomes between the treatment and control groups observed after the intervention can be attributed to the intervention specifically. In other words, the control group is the counterfactual that helps observers understand what would have happened to the treatment group were it not for the intervention.[5]

However, not all financial capability programs are well-suited to be the subject of such rigorous evaluation. Program evaluation—and RCT evaluations in particular—can be time-intensive and expensive, and many programs may not have the operational capacity or client volume to justify participation in such a study. Even well-established programs with large client bases may find it difficult to participate in an evaluation without external support. Further, it is important to consider how delaying or denying service to the control group could impact the control participants. As described in more detail below, to address this issue, the RCT might be conducted as a study of a pilot program or a new service or to perform random assignment

[5] For a more thorough explanation of RCTs see Wholey, Hatry, and Newcomer (2010) *Handbook of Practical Program Evaluation* or Shadish, Cook, and Campbell (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.*

before individuals are even offered the service. Funding supporting these evaluations is limited, which puts more pressure on the studies that are conducted to be well-planned and well-executed.

Participants in the convening agreed that, where feasible, the RCT is the preferred way to measure a program's effectiveness. The convening participants identified key criteria for determining whether to go ahead with a potential evaluation. The first three criteria are factors that cannot easily be addressed in the short run. The final three criteria are likely more amenable to change during the process of exploring the possibility and desirability of an RCT in a specific situation.

Scalability and replicability

RCT studies often have high opportunity costs and actual costs, and research funding should be directed toward studies whose findings will be widely relevant or replicable to other sites.

Organizational capacity and size

The program itself should be well implemented and managed, and have the operational capacity to assist in the implementation of the study. Further, the program should be of sufficient size, or be capable of growing to a size large enough, to support an evaluation (in terms of the number of individuals enrolled in the study and treated). The minimum number of study participants needed depends on the expected size of the effect. The smaller the expected marginal effect of the program, relative to outcomes for the control group, the larger the number of participants needed to reasonably expect the study to detect an impact on clients.

Program stability

Program design can shift in response to staffing changes or funding mandates. It is often quite important that programs demonstrate a track record of stable service delivery, in order to have worked out the kinks in implementation before becoming the subject of a rigorous evaluation. However, an established track record is less critical for interventions that vary a discrete component of an program, or if the intervention itself is less complex or less reliant on people for the delivery of services, such as an RCT involving a new application of technology.

Adequate funding for program operations and evaluation and buy-in from funders

Programs should have sufficient resources to implement the model throughout the evaluation period, while remaining faithful to the research design. This may require additional funding to

increase the number of individuals served or for gathering and entering data, as well as buy-in and support for the study (and its randomized design) from existing funders and board members.

Buy-in and research planning participation from front line staff

Front-line staff who will need to be involved in study implementation on a day-to-day basis need to be fully invested in the study and involved in designing implementation and data collection strategies. This includes the staff responsible for recruiting study participants, assigning participants to treatment and control groups, providing the financial intervention, managing data. As one researcher said, "I won't do a research project with a program without the IT person signing off."

Close working relationship between evaluators and providers

Both parties should see the evaluation as a partnership, have an equal commitment to the fidelity and success of the study, and be willing to listen and contribute to making it work. Both sides should expect a heavy dose of upfront talks during the research design and early implementation phases, as well as frequent ongoing communications once the study is up and running.

3. How to conduct a strong study: Solutions to common challenges

Researchers face a number of challenges in designing and executing RCT evaluations of financial capability programs. The convening participants identified common obstacles to do with research design, research implementation, and data collection and analysis, and they offered ideas for overcoming these challenges.

3.1 Research design

The first set of obstacles researchers and programs face relate to research design. Can the study support an RCT design? What should target enrollment be? How will study participants be randomized? Do new partners need to come aboard for the study to be viable? For questions such as these, the attendees of the convening identified a number of areas that require special attention from researchers and program partners during the study design phase.

Scope of evaluation

One major challenge in research design is determining the appropriate scope of the evaluation—e.g., how many outcomes to measure and how long to follow study participants. Certain outcomes, such as building up savings or improving credit scores, may take months or years to develop or may be costly and difficult to measure. Researchers, practitioners, and funders should be realistic in approaching the tradeoff between the importance of obtaining data on numerous financial behavioral outcomes versus the feasibility of observing change in these outcomes within the scope of the study.

Limited sample size

The biggest challenge facing most RCTs is generating a large enough sample of study participants to measure program effects with statistical precision. Limited resources, difficulty recruiting participants, and participants dropping out before they attend sessions or complete all surveys all hamper evaluators' efforts to reach target sample sizes. A limited sample size may require a research design focused on one or two central outcomes where the largest average effect sizes are anticipated across the full study population. Other outcomes might be considered or explored but not rigorously tested due to smaller average effects and a limited capacity to examine impacts on different subgroups. Many researchers hope that the financial capability field will identify one or two "priority" outcomes that measure basic financial capability across a broad range of populations, contexts, and interventions—analogous to measuring patients' blood pressure and body mass index in health research.

Small, difficult-to-detect changes in behavior

Often, the expected changes in behavioral outcomes are small, which makes them even more difficult to measure. For this reason, it may be prudent to focus on behaviors where single decisions or changes at the margin can have substantial long-term effects, and therefore be easier to detect. Examples of such behaviors are yes/no decisions such as whether a study participant established a savings account or an automatic bill payment, or threshold outcomes such as whether a study participant was able to qualify for a standard credit card (to get away from predatory lending) or a mortgage refinance (to a lower cost loan). One way to improve the likelihood of finding true but relatively small program effects is to enroll a sufficiently large number of study participants. The necessary research sample size to detect impacts of varying sizes should be clearly discussed between evaluators and program staff in the early stages of project feasibility discussions and research design.

Discomfort with denial or delay in service due to randomization

Funders and service providers sometimes balk at the idea of delaying or denying service to control participants. To allay this discomfort, the RCT might be conducted as a study of a relatively straightforward pilot program or a new variant on an existing service. Another option is to perform random assignment before individuals are even offered the service. This approach may work best in a setting where individuals show up for other services – such as a workplace program, tax filing assistance, or applying for a loan – and the treatment is framed as an additional benefit, or where individuals are not the ones initiating contact to begin with. It is

also possible to implement "encouragement" designs, in which the service is available to everyone but is made more readily available to a random subset.

Program and study enrollment selection effects

Certain kinds of interventions rely on individuals to voluntarily participate. In these cases, people who choose to participate in an intervention are likely more concerned about and motivated to address their financial behavior than the general population. While the RCT design should remove any significant differences between the treatment group and the control group, the study participants overall may differ in important ways from the general population. This implies that researchers must be cautious in drawing conclusions for the general population from these results.

Point of randomization

Researchers acknowledged that in some circumstances, randomly assigning individuals to treatment and control groups as they enroll–the "coin-flip" ideal–may be infeasible. Some alternatives to this design that preserve the RCT approach include randomly assigning similar groups of individuals, or enrolling participants to the treatment group on a first-come first-served basis and using the waitlist as a control group, if there is good reason to believe that the order in which people access a particular service is essentially random. Of course, an RCT need not compare an "all or nothing" dichotomy; it is possible to compare interventions with different dosages, designs, or durations.

Model fidelity

Many agreed that once a suitable program is chosen, "the less you can mess with standard business practice, the better," as voiced by one convening attendee. A study will be of greater usefulness if it evaluates what a program does (or would do, in the case of a new approach) during its normal course of business. It is also more likely to be faithfully implemented if it places fewer new requirements on program staff.

Input from program partners

Researchers need to proactively seek staff input on all areas of study design, including determining the study timeline, engaging with potential study participants, measuring changes in financial decision making among their target population, survey design, and data analysis strategies. Staff working directly with clients often provide much needed "reality checks" for the research design: will these plans actually work as intended, or will some idiosyncratic factor

unexpected to researchers but well known among staff on the ground suggest a change in plans? (Target gift cards are not useful in an area with no Target stores; proposed survey question wording may not be familiar to study participants, etc.) Front-line staff knows the study participants and program operations best, and it's important to make use of this knowledge.

Process/implementation studies

Convening participants agreed that impact evaluations can be strengthened by including an implementation study, which is a descriptive account of the organization's goals, strategies, methods, and activities combined with the researchers' account of the roll-out of the RCT evaluation. Process studies provide continuous reporting on the evaluation's progress and help to put the eventual study results in context.

3.2 Research implementation

Even the most elegantly designed RCT can face difficulties when study implementation begins. Practitioners need the most support when transitioning from planning the study to starting the study. Participants had the following insights for moving smoothly from the design to implementation phases.

Prepare, prepare, prepare

The transition from planning to implementation should include some preparatory steps if possible. These include field testing surveys and doing practice runs of recruitment pitches, study consent procedures, randomization tools, and follow-up approaches. Some organizations need to hire additional service providers to handle the increased number of clients resulting from the study. Staff training and oversight specifically for the study, and making sure that staff's questions are answered quickly and easily, are also important at the outset, and for continuing to develop buy-in to the study.

Enlist (and budget for) a research coordinator

Organizations can greatly benefit if they have a dedicated research coordinator whose primary responsibility is to manage the evaluation. The additional work associated with a research study may be too much for program staff to handle alone, and having a research coordinator who can devote significant time to this effort may be the difference between a well and poorly executed study. The coordinator might be employed by either the service provider or by the evaluators,

though there are some advantages to making the coordinator a part of the organization's staff (or choosing an organization that already has a coordinator on staff). A permanent research coordinator gains the experience and authority within the organization to build the capacity to participate in rigorous evaluations in the long term. An on-staff research coordinator is better able to understand the everyday challenges that practitioners face, particularly in difficult tasks like denying or delaying service to control group participants, or asking sensitive and potentially upsetting survey questions. Alternatively, there may be some benefits to having the research coordinator be directly employed by the research organization, including, potentially, employing a coordinator with greater experience in RCT design and implementation. Regardless of who employs the research coordinator, it is helpful if the individual has some training and experience in program evaluation or field research, as well as an understanding of the program's operations.

TABLE 1: RESEARCH COORDINATOR ROLES

Research coordinators can…
Understand both the program's need to keep their primary focus on providing services, and the research staff's requirements to push the evaluation forward and maintain fidelity to the research design.
Translate the relative importance of each side's priorities.
Mediate between the program staff and researchers when the requirements of each may come into conflict.
Minimize the evaluation's disruptive effects on program operations.
Provide oversight and quality control for research implementation and fidelity to evaluation design.
Improve the program's ability to internally track its performance.

Align data collection efforts

In most studies, evaluators are responsible for collecting the data that will be used in the analysis. Evaluators should coordinate with organization staff to minimize duplicative data collection and duplicative contact with participants. Automated randomization and data collection tools that can sync easily with the organization's existing tracking system reduce data

entry burden, ensure fidelity to the research design, and alleviate the stress of randomization, are important components.

Expect bumps along the way

Even with ample preparation, it is entirely normal for things to go wrong. Both evaluators and program staff need to be prepared for pitfalls like lower-than-expected rates of participation in the study, difficulty tracking down participants to receive services, important survey questions being misunderstood or skipped entirely, and IT problems. Good communication between researchers and practitioners is key to moving past these problems. Problems at start-up are less likely to derail an evaluation as long as they are documented and dealt with as soon as they arise.

3.3 Data collection and analysis

Convening participants recommended that data analysis not only be considered as the capstone of the research effort, but be sufficiently incorporated into the planning and research design stages and expand as study data become available. They offered the following advice on the use of data, from the early stages of a study to the final analysis.

Look at data early and often

Researchers first use program data to determine whether the program is a good candidate for evaluation, and to design their own data collection tools to complement the programs' existing intake forms and tracking instruments. Study coordinators can also use participants' data to troubleshoot the randomization, data collection, and data entry processes. Initial responses to financial knowledge and behavior questions can also highlight areas where the follow-up survey may need to be revised.

Don't undervalue intermediate results

Program and research funders are also interested in short-term and intermediate study results to keep the conversation active around ongoing evaluation projects and to help fulfill responsibilities to their stakeholders. These results may also be of interest to practitioners and policymakers. Research designs that incorporate theories of change for improving financial decision making, and logic models for the impact of program interventions, allow evaluators to connect early data to potential long-term outcomes.

Choose data sources deliberately

Administrative data collected by third-party organizations, and other forms of what one funder referred to as "naturally renewable data sources," are increasingly useful to evaluations and to financial capability training more generally, but it cannot be assumed that such data will necessarily be available to researchers. For example, it is becoming standard practice for financial capability programs to review individuals' credit reports as a part of understanding their financial history and how their behaviors affect their ability to obtain credit. Programs or interventions operated by financial institutions or third-party personal financial management firms frequently download information from participants' financial accounts. However, the right to use such data for research purposes must be clearly established in the context of a specific study.[6] Additionally, researchers suggested caution in prioritizing measures that are convenient to collect from administrative sources but that might not reflect appropriate goals or likely outcomes for all participants.

Expect realistic effects

Researchers, practitioners, policy makers, and funders stressed the importance of setting expectations for the evaluation's results, both at the outset and in the data analysis stage. All interventions exist in a complex world with other factors affecting the outcomes of interest, and in many cases the results of an intervention may be statistically significant but small. To avoid mischaracterization of the study's conclusions, all parties should give significant thought to how the results will be framed.

Consult practitioners on interpretation of findings

Researchers should gather input from practitioners and policy makers, including ones not directly involved with the study, to assist the researchers in understanding the meaning and implication of the findings. In some cases, a "negative" result might mask a positive outcome. In one example referenced at the convening, researchers found that debt levels increased among program participants, which initially seemed like a failure to improve financial behavior. When the researchers discussed the results with the program staff, they realized that the result was driven by low-income participants who now had access to credit, which allowed them to smooth

[6] For example, the agreements that financial capability programs have with providers of credit reports may not allow for the data that programs access for clients to be transferred to researchers for analysis.

financial disruptions by borrowing. Looking at these individuals' credit reports, researchers saw that they were using credit responsibly. Flexibility, thoughtfulness, and communication are key to interpreting findings.

4. Implications for practice, policy, and funding

Researchers, practitioners who run financial capability programs, policymakers, and funders can take a number of concrete steps to expand rigorous, data-driven, experimental research on financial capability programs. A number of key themes emerged from conversations during the convening about how these different groups can prime the field for future evaluations in this space.

For **practitioners**, key steps included:

- Encourage a "culture of data" in their operations, and prioritize detailed tracking of program activities and outcomes.
- Use data to isolate key impact metrics—those indicators that programs feel that they are able to move the needle on.

Programs that build this internal data collection and analysis capacity will also be able to document their efforts, refine their approaches, and communicate the value of their work to external stakeholders without a formal, external evaluation. Some programs that have participated in rigorous (RCT) evaluation have found that the heighted attention to data and metrics required by study participation helped them notice less successful elements of their program operations that they would not have otherwise been aware of, leading to meaningful improvements in their service to clients.

For **evaluators**, **funders** and **policymakers**, key steps included:

- Developing strategies and systems to make data collection easier, faster, less expensive, and more efficient. Both researchers and practitioners will benefit from standardized, affordable, and easy-to-use data collection and management programs.

- Working with programs to develop a set of "priority outcomes"— accepted financial capability outcomes shared across evaluation studies.[7]
- Designing standardized data collection efforts around the priority outcomes, once established, and making sure that these outcomes are included in subsequent evaluations.

Practitioners need to be supported financially in their efforts to expand their data collection and analysis capacity. There exists something of a chicken-and-egg problem: Programs need funding to set up data collection systems, but have trouble fundraising without data-driven evidence of their program's impact. Reducing the expense and complexity of installing these systems will help to address this issue. Access to a set of priority outcomes, accepted by researchers, practitioners, and funders, can motivate funders to help programs adopt these practices, and the CFPB[8] and others are currently engaged in this area.

Policymakers are interested in the potential benefits of integrating financial capability strategies into other programs and services. Some attendees suggested supporting return on investment studies into the costs and benefits of such integration. Others were interested in seeing how the programs' impact varies by individuals' initial behavior and attitudes in addition to their demographic characteristics.

Finally, all participants emphasized that external, rigorous evaluations are often expensive and time-consuming. It is important, therefore, to choose evaluations strategically, focusing on those with the most potential to further knowledge in the field.

---

[7] The value and practical challenges associated with establishing and collecting consumer financial outcome metrics are described in detail in the CFPB's November 2013 publication "Empowering Low Income and Economically Vulnerable Consumers: Report on a National Convening," *available at* http://files.consumerfinance.gov/f/201311_cfpb_report_empowering-economically-vulnerable-consumers.pdf.

[8] The CFPB's ongoing work to develop and advance measures of financial well-being and related outcomes is described on page 47 of the 2013 "Financial Literacy Annual Report," *available at* http://www.consumerfinance.gov/reports/financial-literacy-annual-report/, and on page 75 of "Empowering Low Income and Economically Vulnerable Consumers: Report on a National Convening," *available at* http://files.consumerfinance.gov/f/201311_cfpb_report_empowering-economically-vulnerable-consumers.pdf.