

Design and Evaluation Principles for Children's Savings Account Programs

Table of contents

About the Bureau's Children's Savings Account Initiative	2
1. Introduction	3
2. Design principles	5
2.1 Coverage.....	5
2.2 Distribution of program resources.....	6
2.3 Rate of return on savings or investments	6
2.4 Administrative costs per account.....	7
2.5 Sustainability	8
2.6 Scalability	8
3. Developing a theory of change and logic model	10
3.1 Definitions.....	10
3.2 Creating a Theory of Change and Logic Model.....	12
4. Measuring program performance and outcomes	15
4.1 Performance monitoring versus program evaluation	15
4.2 Benefits of program evaluation.....	18
4.3 Sample program evaluation designs.....	19
4.4 Steps to conducting an evaluation	21
5. Conclusion.....	25

About the Bureau's Children's Savings Account Initiative

The Dodd-Frank Wall Street Reform and Consumer Protection Act established the Bureau's functions to include "providing opportunities for consumers to access . . . savings, borrowing, and other services found at mainstream financial institutions." Dodd-Frank established the Office of Community Affairs¹ to provide "information, guidance, and technical assistance regarding the offering and provision of consumer financial products or services to traditionally underserved consumers and communities." While continuing to explore different strategies to further these directives, the Bureau has identified Children's Savings Accounts (CSAs) as a promising way to support economically vulnerable households' access to and engagement with long-term savings and investment accounts. A growing body of research suggests that having even small amounts of savings earmarked for post-secondary education increases a child's likelihood of attending and completing college.² In addition, Bureau research shows that regularly putting money into savings is linked to financial well-being.³

Over the past several years, the Bureau has engaged with CSA programs, researchers, funders, and other stakeholders to build on existing efforts to support the development of resources for the CSA field. The Bureau has created materials and information sharing opportunities to enhance CSA programs' capacity to engage participants, measure outcomes, and learn from one another. Bureau projects have included the development of four [program design guides](#) and a [2018 CSA forum](#) that brought together researchers, representatives from state and local CSA initiatives, and intermediaries and financial institutions affiliated with CSA programs.

This is the first of two guides the Bureau released in October 2020 covering the design and evaluation of Children's Savings Account programs. The guides complement one another and are designed to be read in order. The second guide is titled *Common Metrics for Children's Savings Account Programs*. The two guides build on the Bureau's past work on CSAs and explore how the field might take a more coordinated approach to program design and evaluation. The guides are intended to help programs refine their goals and then align these goals with measures of day-to-day performance and longer-term participant outcomes.

¹ The Office of Community Affairs was formerly the Office of Financial Empowerment.

² Elliott, W., Song, H., & Nam, I. (2013). Small-dollar children's savings accounts and children's college outcomes by income level. *Children and Youth Services Review*, 35, 560–571.

³ Consumer Financial Protection Bureau. (2018). Research Brief: Pathways to financial well-being: The role of financial capability.

1. Introduction

Children’s Savings Account (CSA) programs provide incentives and other support to encourage children and families to contribute to long-term savings or investment accounts dedicated to post-secondary education. CSA programs vary in design, but their primary goal is helping children access and complete post-secondary education by increasing the financial resources, financial capability, and educational expectations of children, parents, and caregivers.

CSA initiatives are becoming more common around the country, with over 80 programs currently in operation or development.⁴ The primary long-term goal of CSA programs is to increase the number of young adults enrolling in and completing post-secondary education.⁵ However, CSA programs vary in their design, implementation, populations served, origin (public or nonprofit), secondary goals, and available resources.

As the Bureau worked with CSA program administrators, researchers, and funders, we identified two key challenges for the field. One is the field’s ability to identify best practices for program design. A second is the lack of a universal set of performance and outcome measures; this lack of standardization has limited the ability of programs – individually and collectively - to define and measure outcomes and impact. It has also limited the capacity for research and evaluation, and ultimately policy development, since measures from one program often cannot be compared to other programs.

Recognizing the potential benefits of a more standardized approach to program design and evaluation, the Bureau has developed these companion guides for CSA programs, researchers, and other stakeholders. This design guide is the first of the two guides, which are intended to be read in order. The second guide is titled *Common Metrics for Children’s Savings Account Programs* (“Common Metrics Guide”).

4 Prosperity Now (2020). *The Movement Reaches New Heights: The State of the Children’s Savings Field 2019*. Washington, DC.

5 Post-secondary education may include college and university programs as well as trade and vocational schools. Each CSA program has guidelines specifying qualified uses of funds.

These guides are intended to help CSA programs develop a framework for identifying and incorporating metrics that enhance the success and sustainability of individual programs, and support evaluation and scalability across the CSA field.

2. Design principles

This section introduces a set of underlying design principles that provide a foundation upon which to build programs that are effective, scalable, and sustainable. These principles may help CSA programs refine their goals and provide direction on how to measure success.

With input from CSA programs, administrators, and researchers across the country, the Bureau identified the following list of core CSA design principles. New CSA programs may consider these principles as they develop their programs, and programs already in operation may reflect on them as they continue to refine their programs and evaluate their progress. It is up to each program to decide how much each principle influences its goals and what types of metrics are most suitable for assessing progress towards those goals. In some cases, the principles may involve tradeoffs with one another. Many of the measures in the *Common Metrics Guide* could be used to assess how well a program is fulfilling each goal. Programs may consider incorporating these principles directly into their stated goals.

2.1 Coverage

Coverage is the extent to which a CSA program reaches the entire population of children eligible for the CSA in its defined service area. In other words, who is included in the program and who is not included? Program design can have dramatic effects on coverage, with universal “opt out” programs generally attaining higher degrees of coverage than “opt in” programs.

Several measures in the second guide may be used to determine a program’s coverage. At a basic level, programs can compare the number of children with active CSA accounts to the overall number of children in the target population. For a more in-depth understanding of coverage, programs may break out CSA participation rates by ZIP code, race and ethnicity, poverty level or socioeconomic status, parents’ educational attainment, or other geographic or demographic characteristics. Any trends in the data may help programs identify geographic areas or subpopulations that are lagging others in terms of their participation, information that could then be used to develop specific strategies to close any gaps.

2.2 Distribution of program resources

Distribution of program resources describes how program funds (e.g., seed deposits, savings matches) are distributed across different groups. CSA stakeholders suggest that many programs want to make sure that benefits adequately reach underserved children and families. What this commitment looks like varies by program. Some CSA programs explicitly prioritize underserved populations, for example, by providing larger seed deposits or savings matches based on certain criteria such as family or neighborhood income. Other programs do not have incentive structures specific to lower-income families but focus their outreach and programming in lower-income neighborhoods. On the other hand, some programs have not set goals around the distribution of program resources. In general, opt-out programs tend to prioritize universal enrollment over a focus on children with the greatest needs, though such programs may have other features designed to advance goals around resource distribution.

Both coverage, described above, and distribution of program resources should be considered together for an accurate assessment of how well a program is meeting the needs of underserved children and families.

Several metrics in the Common Metrics Guide may serve as building blocks for the analysis of how program resources are distributed. For example, programs may break out seed (initial) deposits, program deposits, or other measures by student demographics, neighborhood, school, or other characteristics to help determine what distribution looks like. Taking this a step further, programs may be interested in exploring how their efforts to ensure program deposits reach all students relate to program outcomes (e.g., how much lower-income students benefit from larger seed deposits).

2.3 Rate of return on savings or investments

Rate of return refers to the CSA account holders' earnings from interest or investment—essentially the account's earnings (or losses) excluding any deposits or withdrawals. A rate of return can be measured in dollars or expressed as a percent of an account balance; in the latter case, a rate of return is often annualized. Not all CSA programs prioritize investment growth rate, especially those that use accounts through local banks or credit unions (where returns are limited to interest rates that, at the time of writing, are at historic lows). Instead, these programs may be more concerned with promoting savings behavior and habits rather than emphasizing growth from returns on investment. In CSA programs that use 529 accounts, families have more investment options and can potentially earn higher rates of investment

growth.⁶ These programs may place a higher priority on rates of return, though they may need to factor in downside risk that comes with investing.

Calculating rates of return requires account data on balances, deposits, and withdrawals over time. Programs may also be interested in analyzing if and how rates of return vary based on demographic characteristics such as race and ethnicity or income.

2.4 Administrative costs per account

Administrative costs per account include overhead and the operational expenses required to manage account activities and provide the additional support services to participants. Although all programs incur costs for account management, programs vary in terms of the additional support they offer to participants. CSA programs differ in the extent to which they use high- versus low-touch approaches to engage children and families, and these differences in approach can significantly affect administrative cost per account. High-touch programs inevitably involve higher administrative costs per account than lower-touch programs. However, the additional services offered in high-touch models are often intended to increase child and family engagement and in turn, children’s likelihood of enrolling in post-secondary education. Thus, a full accounting of both costs and benefits is needed to judge a program’s overall efficiency and effectiveness—high-touch models are more expensive to operate, but their benefits may still outweigh the added costs. As the body of evidence about CSA programs grows, programs will be in a better position to compare the costs and benefits of different account structures and engagement strategies.

Input from CSA programs, researchers, and funders indicates that costs are often difficult to compile and analyze, a fact that is particularly important when comparing the costs of different CSA programs. Programs may consider deepening their understanding of their cost structure, accounting for the costs of account maintenance (including enrollment, account management, and reporting) separately from the costs of support services, which are potentially highly variable over time and across programs. There are likely to be economies of scale in program maintenance functions, so that larger programs have lower costs per account than smaller programs.

⁶ Investment accounts such as 529s do present a great risk of loss, especially in the short term due to potential for market volatility. However, 529 plan investment portfolios generally offer age-based and static asset allocations. With an age-based option, the asset allocation is designed to automatically shift away from stocks and move toward more conservative investments as the child gets closer to post-secondary education.

2.5 Sustainability

Sustainability is the likelihood that a CSA program will be able to continue operations for the foreseeable future, serving not only the children who currently have active accounts, but also future children. Sustainability is especially critical for CSA programs given that they are typically intended to serve children for a long time – often 12 to 18 years – before the child enters post-secondary education. Beyond enrollment in post-secondary education, programs may seek to support participants through completion of their post-secondary educational goals.

Sustainability is difficult to measure and highly dependent on program context (e.g., a statewide publicly supported program versus a local program supported by individual donations). Ideally, sustainability measures capture the permanence of different funding sources. For example, a program with an endowment may be considered more sustainable than one dependent on annual grants. Likewise, a state-funded program that is established with a mandated set-aside is likely more sustainable than one where program funding needs to be reauthorized by the legislature on a regular basis.

Sustainability is often linked to the stage of development of a CSA program. Often, newly emerging CSA programs are small and dependent on a specific champion – a single individual or institution that advocates and ensures the success of the program. Over time, programs may work to broaden their base of support to include multiple stakeholders and institutions. Reaching this stage of development is necessary for sustainability and may be more attainable for programs that are more scalable and have lower costs per account.

2.6 Scalability

Scalability is the extent to which the CSA program or its subcomponents are easy to replicate and benefit from economies of scale such that the cost per account declines as the number of participants increases. The primary motivation for adopting a scalable model is the ability to serve more children and families. Many CSA programs are designed initially to provide an account to every child within the program's geographic or jurisdictional reach. Other programs start by serving a smaller number of participants with the intent to expand as the model proves effective and more resources become available.

There are trade-offs between scalability and other program goals. For example, a local program with accounts through a community bank or credit union may be less scalable than one using the state's 529 platform, but it may reach children and families that 529 accounts cannot serve, such as those without a Social Security number. Program engagement strategies also affect scale. A statewide mailing to parents is much more scalable than holding in-person promotional

events at individual schools, but the ultimate value of each approach comes down to costs compared to benefits.

One way of gauging a program's scalability is whether it has the capacity to expand and is using that capacity to serve more children and families. Is the number of children enrolled in the program growing over time? Similarly, is the percentage of children and families actively engaging in program activities increasing?

Another way of viewing scalability is replicability—whether lessons learned from one CSA program may be readily adopted by other programs. Programs that can be replicated with relative ease in another jurisdiction may lead to more children regionally or nationally with accounts, which improves CSA coverage in a broader sense. But not all programs that can be easily replicated are effective, so the degree to which a program is being replicated is an indicator of how the field is growing overall, not a value judgment.

No specific metrics are included in the Common Metrics Guide for scalability, but it is an important policy design feature to consider for program administrators developing new CSA programs. Trade-offs between scalability, and the degree of engagement offered by the program (high touch or low touch), should be made explicitly and purposefully rather than accidentally.

3. Developing a theory of change and logic model

This section discusses the importance of a Theory of Change (TOC)⁷ or Logic Model (LM),⁸ a key step in identifying CSA program goals and outcome measures. A typical first step in developing a program and identifying performance measures is to create a TOC or LM that describes the full sequence of a program’s services and expected outcomes. Individuals and organizations interested in developing programs often start by recognizing an existing need, developing a problem statement and related goals to overcome the need, creating a set of activities to address the need and reach an ultimate goal, and selecting outcome measures based on the prior steps.

3.1 Definitions

The fields of performance measurement and programs evaluation use a variety of similar and sometimes overlapping terms. While some of these terms may seem like jargon, it is important to understand and become familiar with the terms and their usage. Common terms include:

Measures and Metrics. Measures and metrics are numbers used to count or express particular characteristics, such as the number of CSA account holders or the average dollar amount of deposits into CSA accounts. Measures or metrics are observable characteristics or changes that capture progress towards or the achievement of a desired result. Measures or metrics can be further categorized based on how programs use

⁷ *Stanford Center on Philanthropy and Civil Society.* Theory of Change—how we effect change. The overarching set of formal relationships presumed to exist for a defined population, the intended outcomes that are the focus of the organization’s work, and the logic model for producing the intended outcomes. A theory of change should be meaningful to stakeholders, plausible in that it conforms to common sense, doable with available resources, and measurable. www.pacscenter.stanford.edu/publication/logic-models-an-overview.

⁸ *Stanford Center on Philanthropy and Civil Society.* Logic Model—what we do and how. The logically related parts of a program, showing the links between program objectives, program activities (efforts applied coherently and reliably over a sustained time), and expected program outcomes. A logic model makes clear who will be served, what should be accomplished, and specifically how it will be done (i.e., written cause-and-effect statements for a given program design). www.pacscenter.stanford.edu/publication/logic-models-an-overview.

them, with one important distinction being the difference between “outputs” and “outcomes.”

Outputs. The quantity of a program’s actions, such as the number of products created or delivered, number of people served, or activities and services carried out. Outputs are typically reported for defined periods of time (e.g., by month, quarter, year)

Examples: The number of children enrolled in a CSA program during a calendar year, the number of deposits families make into their accounts from September through December, or the number of financial capability classes offered by the program during the school year.

Outcomes. Socially meaningful changes for those served by a program. Outcomes are generally defined in terms of expected changes in participants’ knowledge, skills, attitudes, behavior, conditions, or status. Outcomes are the benefits that participants derive and the changes that result during or after involvement with a program. Outcomes are often expressed using words such as increase, decrease, or reduce. To distinguish between outputs and outcomes, one can think of outputs as concrete actions designed to lead to or result in an outcome. In practice, outputs and outcomes sometimes overlap, and distinguishing between the two can be challenging in some cases.

Examples: Increase in the rate of children and families making deposits into their CSAs, increase in the financial capability of participants, and increases in children’s college aspirations.

Outcomes can also be defined by time periods, with interim outcomes ideally leading to longer-term ones.

Example: Increasing students’ knowledge about the benefits of post-secondary education (interim) may influence long-term increases in participants’ applications to, enrollment in, and completion of post-secondary education.

Benchmarks. A benchmark is data that programs use as a point of comparison to interpret their data. Data from a single program may be difficult to interpret without the context provided by benchmarks. Benchmarks may include data from the same program but from an earlier time period (i.e. to explore trends over time), data from other programs, or a goal or threshold set by a program. In certain cases, programs may be able to identify data from a group of non-participants that can serve as a useful

benchmark. Similarly, a benchmark might also be tied to external information such as the rate of savings of account holders nationwide.

Example: A CSA program uses data on the number of accounts opened during the prior program year as a benchmark to judge its success in promoting account openings in the current year.

3.2 Creating a Theory of Change and Logic Model

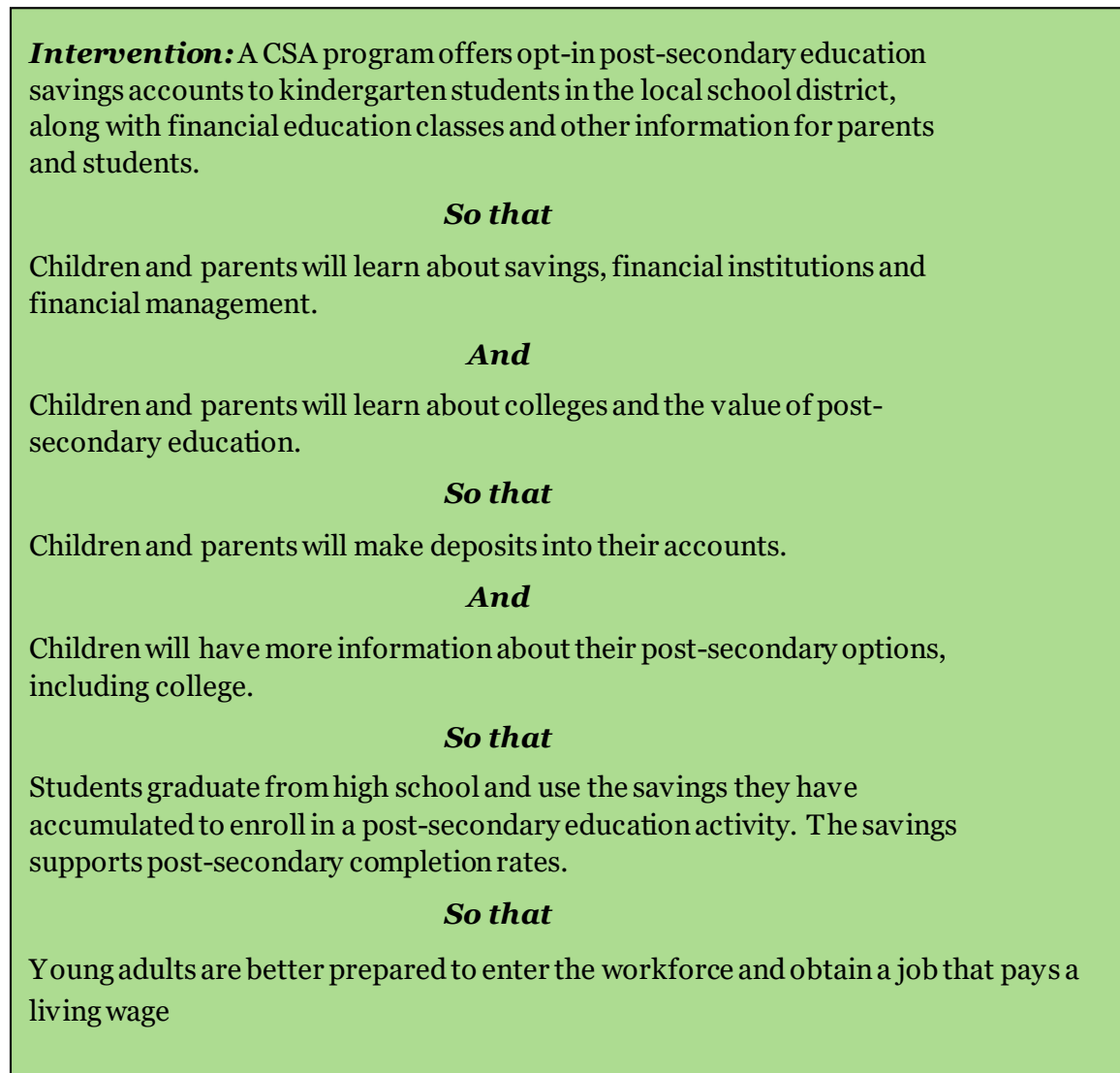
Where do CSA programs begin their effort to identify and incorporate measures or metrics into their program? Before using the Common Metrics Guide, it is important to define the CSA program design and its goals. Then programs can attach specific output and outcome measures to their activities. A first step in this process is the development of a TOC.

Theory of Change

A TOC is a planning tool used in the early development of a program to identify a program's expected outcome(s). In narrative form (as compared to the graphic format of a Logic Model explained next), a TOC helps clarify the program's boundaries given the context in which it operates and identify any gaps in understanding about how the program works.

The process of developing a TOC begins with a situation analysis that articulates the nature and extent of a problem or a challenge, its root causes, and its likely consequences. The next step is to clarify the key issues that the program will address, and the desired outcomes or conditions that the program may affect or reach, as well as noting problems that are outside the program's scope. Then the TOC articulates an outcomes chain showing the assumed cause-effect relationships between program activities and short-term outputs; and short-term and long-term outcomes. Using "if-then" or "so that" statements, the outcomes chain begins with defining the intervention and then proceeds through possible short- and long-term results or outcomes. This process ensures logical thinking around program principles and intended outcomes. Figure 1 is a sample outcomes chain for a CSA program.

FIGURE 1: THEORY OF CHANGE SAMPLE

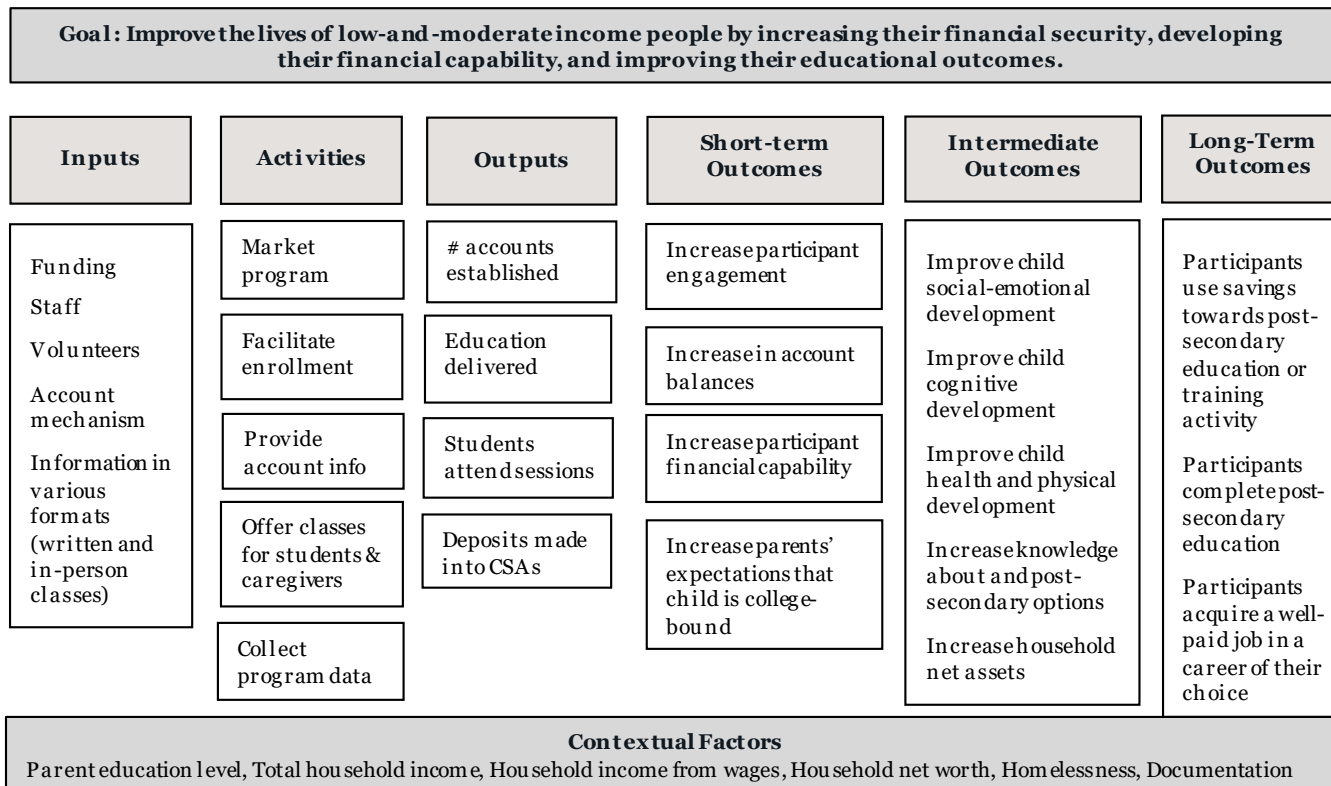


Laid out this way, an outcomes chain set of statements that comprise the theory of change can then help pave the way for development of the pieces of the logic model.

Logic Model

A LM illustrates the inputs and activities of a program, planned outputs, and resulting outcomes. A LM illustrates the pathways or sequences of activities (inputs) that lead to those interim and longer-term outcomes. Figure 2 shows a standard format for a logic model, based on a hypothetical CSA program.

FIGURE 2: LOGIC MODEL SAMPLE



A logic model like the one depicted here tells the intended story of the program, noting a set of activities and inputs, demonstrating how those activities and inputs lead to outputs, and then to short, intermediate, and long-term outcomes. Once a program describes the intended change and the steps needed to realize that change, appropriate outcome measures can be identified. In practice, intermediate and long-term educational outcomes often require a substantial period of time to observe, especially when accounts are established at birth or a young age. Because of the substantial length of engagement with a CSA program by children and families, short-term and intermediate outcomes can be used as measuring sticks or signposts along the way to help determine program effectiveness and to adjust as necessary.

4. Measuring program performance and outcomes

This section discusses the differences between performance monitoring and program evaluation and describes potential approaches to evaluation to orient readers to key aspects of conducting a program evaluation.

CSA programs have different goals and capacities in terms of how they monitor their performance and evaluate participant outcomes. In addition, programs face different measurement and reporting obligations, often related to funder requirements. How CSA programs monitor their performance and evaluate results looks different across programs. Even if two programs adopt the exact same set of measures (such as those proposed in the Common Metrics Guide), they may collect and analyze the measures in different ways and for different purposes.

Reflecting the diversity in the ways programs collect and use data, terms like “evaluation” may refer to a range of activities. This section describes some of the key types of activities that fall under program “monitoring” or “evaluation.” At one end of the spectrum is an evaluation led by a specialized group of researchers who randomize participants into control and treatment groups, but far more common are the day-to-day data collection activities that programs use to run their programs and report results. Metrics proposed in the Common Metrics Guide may be used in different types of assessment or evaluation projects. For example, the same metric may be used both for performance monitoring and program evaluation, the first distinction in this section, described below.

Note: This section is not intended to capture the entirety of the Bureau’s evaluation framework; rather, it aims to present the elements most relevant and applicable to CSA programs.

4.1 Performance monitoring versus program evaluation

Performance monitoring is the use of data to inform the day-to-day or ongoing operations of a program. For example, a CSA program may regularly review the number of new accounts opened at different schools to identify trends and possibly intervene at schools with low

numbers of account openings. Alternatively, a program may seek out schools with high numbers of account openings to identify promising strategies to replicate elsewhere. Using data for these types of purposes is sometimes called performance “management” to emphasize that program administrators actively use data in the course of making decisions. Program managers and other stakeholders can use performance monitoring to enhance the design, delivery, and quality of services of the program. Performance monitoring may also be useful for testing possible program changes; programs may pilot changes on a limited scale before deciding whether to roll them out more widely.

Programs often develop performance monitoring systems to track how well they are meeting certain goals or benchmarks, paying attention to trends over time. Frequently a program will have systems in place to assess specific measures at regular intervals, such as quarterly or semiannual reviews of account openings and balances. Funders often require a specific set of performance reporting from programs.

Example of CSA Performance Monitoring: A newly established CSA program sets a goal of 200 students opening CSA accounts in the first year. To monitor progress toward this goal, the program plans to compile account information quarterly. After the first quarter, staff find that only 38 accounts have been opened, most of which were opened at a single school. Staff then examine possible reasons why account openings are lagging at other schools and potential remedies. Staff adjust program strategies to help the program get back on track toward the target. The staff members also decide to begin reviewing account openings monthly so they can make sure their new strategies are working as intended. A statewide CSA program could take a similar approach, but at a larger scale. Instead of looking at performance school-by-school, a statewide program may look at performance across school districts or other units of geography.

Program evaluation is focused on measuring outcomes and impact. In the most rigorous or formal sense of evaluation, evaluators use methods like randomized controlled trials (RCTs; discussed below) in order to demonstrate that a program caused the measured results. From the adage “correlation is not causation,” this highly formal type of evaluation is designed to identify how a particular intervention (or set of interventions) impact the outcome of interest by randomly assigning people to either receive or not receive the intervention. For example, an evaluator may test how much a CSA program incentive structure caused an increase in deposit amounts, versus how much participants would have deposited without the incentive. Results from program evaluation can inform the strategies of other CSA programs, whereas performance monitoring is generally more inward facing at the operations of the specific CSA program.

Example of an RCT: Evaluators work with a local CSA program and the school district to randomly assign some schools to integrate college workshops into their CSA programming, and others not to do so. The evaluators survey students one year later to determine whether the workshops influence students' expectations around post-secondary education. Similarly, a statewide program could vary certain aspects of program like incentive amounts or marketing messages across school districts, track outcomes over time, and use this data to estimate what effects the variations had.

For many programs, other forms of evaluation may suffice, as RCTs require significant resources and considerable coordination (e.g., an evaluator may require certain schools or classrooms to modify specific practices to fit the needs of the study). Since so many CSA programs were established recently, evaluation often benefits from a certain degree of program maturity or stability, something programs that are still experimenting and expanding rapidly may not have attained yet. Evaluating a program in flux may yield results that are not applicable once the program evolves and looks different. Similarly, a program may need attain a certain degree of scale before an evaluation so any statistical tests have a sufficient sample size.

Example of an evaluation without random assignment: An opt-in CSA program tracks participants' Free Application for Federal Student Aid (FAFSA) completion rate because this is an important step in preparing to attend post-secondary education. In this case, the program does not have data on nonparticipants to assess whether they completed the FAFSA at a different rate than participants. Although this program cannot compare FAFSA completion rates between participants and nonparticipants, the FAFSA completion data it does have is valuable. For instance, the program could use this data to identify schools where FAFSA completion rates for participants appear higher or lower than others and then explore possible reasons for these differences.

In addition to program evaluation, CSA programs may also be familiar with “process” or “implementation” evaluation. Process evaluations examine how well a program is adhering to its logic model and looks at the relationships between inputs, activities, outputs, and outcomes. Whereas program evaluation looks at outcomes, process evaluation examines how different parts of a program fit together and lead to specific outcomes. For example: A program finds that certain materials are not reaching students and parents, which may indicate the need to change their distribution processes to make the materials more accessible.

The purpose of a program evaluation is to measure outcomes, while a process evaluation examines the chain of inputs, activities, and outputs that contribute to program outcomes. A process evaluation may involve more qualitative data (e.g., interviews) than a program evaluation. Overall, while programs are interested in and judged according to their outcomes, it is important to establish ways to monitor and evaluate implementation of the program. Both performance monitoring and process evaluations have a role to play in this regard.

The following sections present some of the different approaches to program evaluation. The information is intended to demystify what evaluation is and to encourage programs to think long-term about what types of research or evaluation they may be interested in. Programs may explore potential partnerships with universities or other entities to discuss whether their CSA program is a good fit for evaluation now or in the future. Programs may make certain design decisions in the nearer term to set the stage for later evaluation.

4.2 Benefits of program evaluation

Programs that have been in operation for several years, have accomplished a specific set of goals, or are interested in knowing more about their impact, may be interested in program evaluation. Evaluations can provide several direct benefits to a program and its participants. Not only can an evaluation show whether a program produces positive outcomes (e.g., program participants apply for and start post-secondary education, use savings accrued to pay for post-secondary education), but when paired with a process evaluation they can illuminate how the program achieved results.

Below are some of the ways that evaluations can positively affect CSA programs and ultimately the children and families they serve.

- Program evaluations can identify ways programs can **improve their services, scale effective practices, and ultimately produce better outcomes for participants.** Evaluation establishes whether a program produces positive outcomes and ideally, exactly how it does so. Learning that a program does not produce certain envisioned results may be just as valuable as learning that it does. Often just the process of participating in an evaluation helps programs deepen their understanding of their services and identify possible improvements.
- Evaluations can **help CSA programs, funders, and researchers make decisions about how they allocate their resources.** An evaluation may help a CSA program identify which activities are more engaging or effective than others. The program may use this information to reallocate its funding or personnel towards higher value activities. An evaluation may also help a program demonstrate its effectiveness and make the case for continuing or even scaling up based on its past performance. Funders in both the public and private sectors may use evidence from evaluations to make decisions about which programs to support and which strategies to help scale up.
- In addition to benefiting the evaluated program directly, evaluation findings can also **provide benefits to the asset building field more generally including policy makers, funders, and other CSA programs.** Evaluation findings may garner

support for the development or expansion of similar programs in other communities. Similarly, positive outcomes may prompt other localities operating different programs to adapt parts of their programs given the evaluation findings. Less positive or conclusive outcomes can also be instructive for other programs considering how to structure their CSA interventions.

4.3 Sample program evaluation designs

This section presents three common ways of conducting programmatic impact evaluations. Other approaches exist, and even the three approaches here may be modified in a variety of ways to meet the needs and interests of a specific CSA program. In general, tracking outcomes without a comparison group is the easiest design to implement but provides the least conclusive findings. A randomized controlled trial is the most resource intensive but does the best job of demonstrating causality—that the CSA program caused observed outcomes. Quasi-experiments generally fall somewhere in the middle in terms of difficulty to implement and the strength of their findings.

Tracking outcomes without a comparison group

In this approach, a CSA program collects data at multiple points in time, generally before the participant enters a program and then at regular intervals over time. The evaluator takes a snapshot of participants at different points in time and analyzes the data to see what changes occur. The evaluator does not have data on a comparison group, only on CSA participants. Data may come from two points in time (i.e. a “pre/post-test”) or from multiple intervals.

Example of tracking outcomes with no comparison group: An opt-in CSA program’s sign-up form asks parents how certain they are that their child will complete post-secondary education (“baseline”). The parents complete a follow-up survey three years later asking the same question. The program looks at how parents’ responses changed from baseline to follow-up.

This approach cannot rule out that the changes from one period to another were due to factors outside of the program or that the changes would have occurred regardless of the program. Evidence from this type of analysis is suggestive but by no means definitive.

Tracking outcomes without a comparison group is often a CSA program’s first foray into program evaluation. Because this type of evaluation does not use a comparison group or require significant modifications to program delivery, it is appropriate for relatively new interventions or programs.

Quasi-experiment with a comparison group

A quasi-experiment compares outcomes of a treatment group (people who participated in the program) to outcomes of a comparison group (people who did not participate in the program) to estimate a program's effects on participant outcomes. In this case, participants are not randomly assigned to the treatment or comparison groups, but instead the evaluator must identify an appropriate group of non-participants to serve as a comparison group or use available data to construct a comparison group. The evaluator uses statistical techniques to control for differences between the two groups and thereby estimate the effects of the CSA program. Evaluators using quasi-experiments must consider the possibility that unmeasured variables may bias the results. If the comparison group differs from the treatment group in unmeasured ways, then the results of an evaluation may reflect those differences rather than the CSA program's impact. There are a variety of approaches to conducting quasi-experiments.

Examples of quasi-experiments: A state rolls out a CSA program in certain areas of the state before others. Students where the CSA operates are the treatment group. The evaluator then identifies a comparison group of students from elsewhere in the state who do not yet have access to the program but look like the treatment group in terms of household income, academic achievement, and other relevant characteristics. The evaluator compares outcomes between the two groups to estimate the CSA program's effects. Another approach is to use cohorts of students prior to a CSA's rollout as a comparison group, observing both this group and the treatment group over time to estimate what effect the CSA has. Still, the challenge exists in determining whether the CSA program was responsible for differences between the treatment and comparison groups, or whether they are attributable to something else.

Randomized controlled trial (RCT)

An RCT is typically viewed as the most rigorous way of establishing that a program caused specific outcomes. CSA participants are randomly assigned, as if by lottery, either to a treatment group, which has access to the intervention, or to a "control group," which does not. Random assignment helps make the two groups as similar as possible, except that members of one have access to CSA intervention and the others do not.⁹ Therefore, differences in outcomes between these groups (e.g., different rates of post-secondary enrollment) can be attributed to the intervention. RCTs are often the most expensive and time-consuming type of evaluation to

⁹ In practice, group "crossover" may occur, where some members of the control group access the intervention and some members of the treatment group do not. In addition, groups in an RCT may differ in certain respects due to chance; this is especially true for smaller sample sizes.

conduct, but they are the most reliable indication of whether the interventions tested had a causal relationship with the resulting outcomes. Because of this, RCTs are often considered the “gold standard” approach to evaluation.

Importantly, like other forms of evaluation, an RCT may look at certain subcomponents of a CSA, not necessarily the program as a whole. For example, while both the treatment and control groups have access to the CSA, the treatment group may have access to something extra that the control group does not. In that case, the RCT is evaluating effects of just that extra offering, not the program in its entirety. RCTs may raise concerns about denying services to the control group, an issue that can sometimes be avoided through creative design. For example, participants in the control group may receive the same service, just after a waiting period.

RCT in practice: In 2007, SEED Oklahoma Kids (SEED OK) randomly selected newborns in Oklahoma to participate in a study of the effects of a CSA based on a statewide 529 plan.¹⁰ About 2,700 mothers completed a baseline survey, after which their newborns were randomly assigned to a treatment or comparison group. CSAs were automatically opened for the treatment group. The treatment group also received a \$1,000 initial CSA deposit, and low-income participants were eligible for an additional savings match. SEED OK tracked outcomes over time for both groups and has documented a range of positive financial outcomes attributable to the intervention. The results found positive outcomes around children’s behavioral and socio-emotional development, mothers’ educational expectations for their children, and parenting practices. If a CSA program can enroll enough students in an evaluation (something an evaluator can assess), local programs can take a similar approach—assigning some students or schools to receive a particular CSA intervention and others not.

4.4 Steps to conducting an evaluation

Whether done internally or by an external evaluator, conducting an evaluation involves several steps. Early steps in the process generally include determining the scope of what the evaluation will cover, specifying research questions, determining the type of evaluation and finally, specifying the expected final products such as reports, briefs, and other written materials. While some organizations may have in-house capacity to conduct an evaluation, programs may contract or partner with an external evaluator. Key steps in conducting an evaluation are described below.

¹⁰ Center for Social Development. “SEED for Oklahoma Kids (SEED OK).” Accessed June 1, 2020: csd.wustl.edu/items/seed-for-oklahoma-kids-seed-ok

Determine purpose and scope

An initial step for an evaluation is to clarify the purpose and scope. Clearly articulating what will be learned from the evaluation will also help to gain stakeholder support for the evaluation activities, as stakeholders will understand the purpose and potential benefits to the organization from the evaluation. During this process, it is useful to refer continually to the program's logic model to ensure that all aspects of the program are considered and to identify appropriate metrics for the evaluation.

Identify research questions and evaluation type

Once the goals of the evaluation and the program components that will be evaluated have been determined, a next step is to craft the key research questions that will guide the evaluation. Strong research questions are rooted in a firm understanding of the program's activities and past programs' demonstrated impacts. Having realistic and specific expectations will help in the design of research questions that are specific, measurable, reasonable, appropriate, and answerable. Strong research questions share the following characteristics:

Specific and measurable. The research questions should identify specific program components or outcomes. This level of detail also makes it easier to identify suitable metrics.

Reasonable and appropriate. Research questions that are reasonable examine outcomes that can realistically be achieved given the time available and expected level of effort of the program intervention.

Answerable. Research questions must be answerable. There are many reasons why a research question may *not* be answerable; data may not exist to answer the question, or the outcome of interest may not be sufficiently defined.

Clarifying the research questions and identifying the outcomes to measure lead to the determination of the type of evaluation that can be conducted and the timing for the evaluation process. These steps are all part of developing the Evaluation Plan. To determine the type of evaluation to be conducted, it is useful to consider two factors:

Determining which method will best answer the research questions. Some questions, such as those about participants' experiences opening and managing their CSA accounts, may best be answered with a process evaluation, whereas questions about the effectiveness of the program are likely to be best answered with a program evaluation such as a randomized controlled trial. Determining the best research methods requires

identifying which research questions are most important and weighing the added costs of more intensive evaluation designs against other priorities.

Organizational capacity to participate in the evaluation. The feasibility of carrying out a specific type of evaluation must be discussed amongst program managers and other stakeholders and should then be documented in an evaluation plan. Even when contracting with an external evaluator, the evaluation will require a significant amount of time from staff to provide information and data to the evaluator, and it is important to keep this in mind.

Data Sources and Collection. Programs may already collect a variety of data points in the normal course of business. Often, evaluation will require a program to make changes to its data points and collection processes, and sometimes implement new data collection strategies such as surveys. For evaluations that include comparison groups, one potential challenge is ensuring that the same data is available for the comparison or control group as the treatment group; data collected in the course of running a program may be unavailable for non-participants. Potential data sources include surveys, agency administrative records, state or federal administrative records, interviews, focus groups, and document reviews.

External evaluator

While some organizations may have the capacity to conduct an evaluation internally, other programs may choose to work with an external evaluator for a variety of reasons. Evaluations require specialized expertise and knowledge that program staff may not possess. Whether working with an external evaluator or not, programs likely need significant staff time devoted to the evaluation. In addition, an external perspective can help ensure the results are objective and independent. Finally, an evaluator with the requisite expertise can develop reports and other written materials that will present results in a way that can be most useful to the program.

In looking for and selecting an evaluator, CSA stakeholders suggested that programs assess whether the evaluator has (1) experience implementing the proposed evaluation design for CSAs or similar type of programs; (2) senior staff with a post-graduate education and some years of experience that demonstrate the technical skills necessary to implement the study; (3) the capacity and resources to facilitate meetings, data collection, data analysis, and report writing; and (4) other specialized knowledge that is important to understanding the program.

Reporting on findings

The reports that an evaluator produces are the culmination of the evaluation activities; they convey the evaluation's findings to relevant stakeholders, including the sponsoring organization, the larger community, and funders. Ideally, evaluation findings help improve program operations and support program sustainability. Findings may also help inform the development and implementation of other CSA programs.

Not all reports come at the end of the evaluation, however. Evaluations may include earlier feedback to help shape program design and implementation. Others include interim reports that typically share findings about the intervention. Final reports serve as the official record of the evaluation and can present process and outcome findings. When contracting with an evaluator, CSA stakeholders suggested it may be useful for programs to define upfront the types of reports they expect and their intended audiences. Similarly, specifying early on whether the program will need both a technical analysis of the research and a more user-friendly description of the findings not steeped in research language may be helpful both to the evaluator and final audiences.

5. Conclusion

The information provided in this guide makes three basic contributions to the CSA field. First, it offers some key principles that programs may want to consider when designing or expanding their CSA programs. The coverage, scalability and sustainability of the program are fundamental to helping more children and families focus on and achieve the longer-term goal of post-secondary education. The programmatic cost per account is variable based on program design, size and depth of services, but still an important factor for program efficiency and effectiveness. The rate of return on investment influences the accumulation of funds in the CSA but the rate of return may be a secondary consideration for programs that are focused on the accessibility of accounts by participants.

Second, this guide offers important information on how to develop a theory of change, which can provide a foundation for program design and a basic set of principles that help to focus the program's intent and purpose. From the theory of change, programs can develop a logic model that provides a nuts and bolts approach to channeling resources to the appropriate inputs leading to outputs and the desired short, medium and longer-term outcomes.

Third, this guide offers a description of various evaluation options that a program should consider as it decides how to measure the effects and impact of the programmatic offerings on the children and families it is attempting to serve.

This guide is also intended to provide an entry point to consider which metrics to use to measure efficiency and effectiveness over time. The companion *Common Metrics for Children's Savings Account Programs* provides a roadmap for selecting the measures a program will use based on its goals, and guidance on the data and information needed to utilize the metrics as effective measures of program and participant performance.

Finally, these two guides offer a foundation to establishing a set of common metrics that can be adopted and used across the child savings account field. Utilizing a set of common metrics will help programs better measure and report on the collective impact CSA programs have on increasing post-secondary education opportunities across the country, especially for economically vulnerable children and families.