

Introduction to program evaluation and its research cycle

This paper introduces the basics of program evaluation. It explains how evaluation can enhance financial literacy programs by examining how they operate, by identifying their strengths and weaknesses and by determining whether they achieve their intended goals. Various forms of evaluation can be applied at different stages of a program. The paper explains when, how and why to conduct an evaluation by discussing key concepts such as the theory of change and the evaluation cycle.

Author

Sandra Mauser
OeNB, Financial Literacy and Culture
Division
sandramausser@hotmail.com



Why evaluate?

Evaluating financial literacy programs can show if they really help people improve financial knowledge and behavior. Evaluation supports program improvement through learning and reflection, and it strengthens accountability by showing whether public money is being used with the intended impact.



Before evaluation begins

Before starting an evaluation, it is important to check if the program is ready to be evaluated. An evaluability assessment helps clarify goals, available data and whether the evaluation can lead to useful results. It also identifies the possible scope of evaluation.



Evaluation is a cycle

Program evaluation is an ongoing learning process. It begins with planning, continues with data collection and analysis and feeds results back into improving the program. Each cycle of evaluation can help to make the program stronger and more effective.

Abstract

Financial education programs aim to improve people's financial well-being. To determine whether this goal and associated outcomes are achieved, it is essential to evaluate how effective these programs are. Program evaluation is thus critical for assessing performance, identifying strengths and weaknesses and guiding improvements. This paper explores the research cycle of program evaluation, covering key phases such as planning, implementation, analysis and communication. It discusses the role of formative and summative evaluation, the importance of stakeholder involvement and the methodological choices required for rigorous assessment. Furthermore, it highlights challenges in evaluation, including biases in interpretation and dissemination. In addition, it underscores the necessity of a structured communication plan to ensure that findings are conveyed effectively to relevant audiences. By understanding the evaluation cycle and its implications, decision-makers can enhance the accountability and impact of financial education programs, thus fostering better financial outcomes for individuals and communities.

Financial education programs are meant to help people strengthen their financial competencies and, as a result, improve their financial lives. A lot of institutional, political, national and even international effort is put into designing and implementing such educational programs. This costs time and money, which raises the question whether the resources are allocated properly. In other words, whether the implemented programs operate as intended and lead to the desired outcomes. This is where program evaluation comes in.¹ Berry and Sloper (2016, p. 5) use the following analogy to describe the purpose of program evaluation: *“A financial education program without evaluation is similar to an explorer without a compass. Without a compass, an explorer is not able to decide whether he or she is on the right track. Without an evaluation, the educator is not able to decide whether the financial program is producing successful results and meeting the audience's needs.”*

1 Program evaluation and its purpose

Program evaluation helps decision-makers set common goals for financial education programs, assess the latter's effectiveness, measure performance and identify challenges and guiding principles. Evaluation results can be used to improve a program's structure, administration or funding. Such results can also help respond to political pressure regarding the program's eligibility (Berry and Sloper, 2016; Rossi et al., 2019). Priority number one for evaluators when conducting an evaluation is finding out what the purpose of the evaluation is: who wants the evaluation, what exactly do they want and why (Rossi et al., 2019; Chambers et al., 2009). For evaluators, it is not always easy to get answers to these questions; they often need to examine the situation carefully. Sometimes, there are *underlying hidden agendas* to program evaluation. For instance, an evaluation might be conducted to strengthen public relations (PR), promote the program among funders and politicians or to provide rationale for decisions that have already been made. To some extent, (almost) all evaluations involve some political and PR objectives. As long as such objectives do not define the main purpose of the evaluation, evaluation integrity is not compromised (Rossi et al., 2019).

¹ Although we are talking about *financial education programs* in this paper, it should be noted that evaluations are conducted for a variety of interventions (programs as well as policies) in different areas of research. Typically, the interventions of interest are implemented to address social problems, e.g. poverty, inequality, unemployment, racism and problems related to healthcare or education. Often these problems stem from underlying social, economic or political factors and can affect a significant number of people within a society. Social problems require public attention and purposeful, well-organized interventions, which also include thorough evaluation (Rossi et al., 2019).

Every financial education program can be regarded as a social framework comprising roles and activities of various groups and individuals who are involved and/or interested in the program. They are connected by a set of social and potentially political relationships. These groups and individuals are called the *stakeholders* of a program. Not only do stakeholders define the underlying social problem to be tackled and the goals of a program, but they also (indirectly) define the purpose of evaluation and influence the way in which an evaluation will be designed (Rossi et al., 2019). Stakeholders broadly fall into three categories, namely (1) people involved in program operations (e.g. program planners and decision-makers, staff, administrators, oversight boards), (2) people served or affected by the program (i.e. recipients) and (3) people intended to use the evaluation findings (e.g. program and evaluation funders, political decision-makers, legislative committees, organizational leads, research community). It can be fruitful for the evaluation process and also for the interpretation of findings to actively include these stakeholder groups' unique and diverse perspectives (Berry and Sloper, 2016).

In a *participatory research approach*, evaluators involve stakeholders directly in the evaluation process (Yoong et al., 2013). Evaluation can be participatory in various ways and at different points in time of the evaluation process. For instance, it is possible to allow program staff, recipients or other stakeholders to provide input for defining both research questions and outcomes to be examined, for interpreting findings and for communicating results to an interested audience. What needs to be carefully considered is how and to which extent evaluators include stakeholders' expertise and experiences in the evaluation. To ensure that the evaluation is not entirely dictated by stakeholders' subjectivity, evaluators are responsible for balancing this input against the objectivity of the process (Yoong et al., 2013).

In this paper, we will focus on two main objectives of program evaluation, namely program improvement and program accountability. These two functional objectives correspond to formative and summative evaluation, respectively (Berry and Sloper, 2016; Rossi et al., 2019). *Formative evaluation* helps determine whether the program's processes and activities are implemented as planned. Identifying the strengths as well as barriers and weaknesses of the program provides information on its quality and guides *program improvement*. Hence, formative evaluation focuses on processes to make adjustments as needed (Berry and Sloper, 2016; Rossi et al., 2019). *Summative evaluation* helps identify outcomes associated with the program and examine whether these outcomes are a result of the program. A financial education program is expected to make a beneficial contribution to the financial lives of individuals. By focusing on the results of the program, summative evaluation can help determine whether these expectations are met. Hence, summative evaluation investigates *program accountability*, providing a summary judgment of the performance of the program (Berry and Sloper, 2016; Rossi et al., 2019).

Box 1

Formative evaluation

To examine components of program implementation quality, formative evaluation could address the following questions (Berry and Sloper, 2016):

- Has the process of delivering information via the financial education program been implemented as intended?
- Can the entire content be delivered to the recipients as intended?
- Are the program's materials used in an engaging manner?
- Do educators and recipients behave in an interactive way?
- Do the targeted recipients use the offered program?

Summative evaluation

To examine recipients' experiences with the program and the outcomes of participation, summative evaluation could address the following questions (Berry and Sloper, 2016):

- Can intended changes in outcomes be observed for recipients having participated in the financial education program?
- Are there observable differences in outcomes between recipients of the program and nonrecipients?
- Are recipients satisfied with the program as provided?
- Do recipients perceive advantages of having participated in the program?
- Do recipients report long-term benefits after a specific time following the program?

Evaluation objectives are often related to a program's *stage of maturity*. With programs in an early stage of implementation, attention should be paid to program stakeholder needs, invested resources and the key activities of a program (formative evaluation). With programs that have been implemented for a while, evaluation objectives might be to understand the short-, medium- and long-term outcomes of a program (summative evaluation) (Yoong et al., 2013; Berry and Sloper, 2016). Program evaluation does not need to concentrate exclusively on either formative or summative objectives. Consistency and quality of program implementation play an important role especially when program outcomes are investigated. The features of implementation will influence how and to which extent the recipients benefit from the program. Hence, evidence from formative evaluation can provide valuable inputs for exploring program impact during summative evaluation more thoroughly (Berry and Sloper, 2016). This is why the functional objectives of formative and summative evaluation are part of a more comprehensive categorization of evaluation types, namely process and impact evaluation.² The purpose of process and impact evaluation is to “*help decision-makers understand whether a program delivered what was promised and how well it performed relative to other programs in terms of benefits to participants*” (Yoong et al., 2013, p. 26). Ultimately, the gained information should help decide on future steps regarding the program (Chambers et al., 2009).

Process evaluation examines the development and implementation of a financial education program (Yoong et al., 2013; Rossi et al., 2019). The goal is to understand how well the program has been executed and how its processes work. It is meant to deliver information on the program's relevance, efficiency and effectiveness. The questions answered by process evaluation can be grouped into three main areas: (1) implementation/operations, (2) outputs and (3) appropriateness/acceptability. Process evaluation focuses on lessons learned and tends to be formative, but it can also include summative objectives (see example questions in box 2). Identifying issues and problems during program implementation is especially important as the program is being rolled out. Conducting process evaluation at an early stage of maturity therefore makes it easier to adjust the program. Stakeholders typically interested in process evaluation are program planners, administrators, oversight boards and funders (Rossi et al., 2019). For them, evaluation insights need to be concrete, timely and immediately applicable.

Impact evaluation examines whether the financial education program is achieving the predefined goals and objectives (Yoong et al., 2013; Rossi et al., 2019). It helps understand which observed outcomes can and

² Examining the literature on evaluation research reveals that the distinction between formative vs. summative evaluation and process vs. impact evaluation is not always clear cut. While these categorizations describe distinct yet complementary roles in understanding programs, they often overlap, and the direction of inclusion remains ambiguous. In this paper, we follow the understanding that formative and summative evaluation objectives define the functional level of evaluation – i.e. which questions to ask – whereas process and impact evaluation reflect the broader purpose of evaluation, placing it in the context of underlying social (and political) goals.

cannot be directly and credibly attributed to the program. In other words, impact evaluation measures the *causal effect* of a program on the observed outcomes, ruling out that any other factors could have caused the effects. The desired effects typically refer to behavioral changes in recipients' financial decision-making. They can, however, also include changes in knowledge, skills and attitudes potentially underlying behavioral changes. The investigated effects can be short-, medium- or long-term. Typically, the long-term effects of a program are referred to as its *impact*. For evaluators it is not only desirable to measure the magnitude of impact but also to explore how this impact was achieved by the program. Stakeholders typically interested in impact evaluation are major decision-makers of program oversight, e.g. program funders, political decision-makers, legislative committees and organizational leads (Rossi et al., 2019).

Box 2

Process evaluation

To examine a program's development and implementation, process evaluation could address the following questions (Yoong et al., 2013; Rossi et al., 2019):

- Were the program's goals and objectives defined appropriately for the given context? (formative)
- Was the program implemented as intended? How can implementation be optimized? (formative)
- Did the program reach the intended recipients? If applicable, why are some eligible people not reached? (formative)
- Are recipients satisfied with the program as provided? (summative)
- Of the people in the program, does a sufficient number complete the program? (formative)
- Did the program have any unforeseen effects? (summative)

Impact evaluation

To examine the causal effect of a program, impact evaluation could address the following questions (Rossi et al., 2019):

- Are the program's goals and objectives being achieved? (summative)
- Are there observable tendencies as to the desired outcomes? (summative)
- Does the program have any adverse effects on the recipients? If yes, which? (summative)
- Does the program have different effects on different groups of recipients? (summative)
- Does the program alleviate the problem it is supposed to address? If yes, to what extent? (summative)

Besides program improvement and accountability, program evaluation also contributes – with its findings – to academic research, *generating knowledge* for the community interested in evaluation studies (Rossi et al., 2019). Since the research field of program evaluation is still developing, good evidence for understanding the effects of financial education programs is still sparse (Yoong et al., 2013). Yoong et al. (2013) point out two underlying challenges: (1) standardized and common benchmark measures have yet to be established to allow comparison between different programs and (2) the existing methodologies differ substantially in their quality and nature. Therefore, any additional knowledge on the evaluation process itself and on the causal relationship between a program and its outcomes helps researchers establish more rigorous program evaluation benchmarks, from which also practitioners and policymakers can benefit.³

³ Addressing the lack of coherent evaluation evidence on financial education, Lorenz et al. (2025b) provide a thematic review of the financial literacy evaluation landscape in OECD countries from 2010 to 2024 that analyzes 68 (quasi-)experimental and peer-reviewed studies. The review identifies key patterns across study contexts, intervention types, outcome measures, results and methodological approaches.

2 Program theory and evaluability assessment

Evaluation can be regarded as an integral part of financial educational programs. To benefit optimally from an evaluation, the evaluation plan should be drawn at the beginning of the program design and implementation. In other words, the evaluation should be *prospective*. Once a program has started, an evaluation can still be conducted *retrospectively*. Yet, without early planning, the information gathered might be limited (Berry and Sloper, 2016; Yoong et al., 2013). Planning an evaluation from the start enhances the quality of evaluation. This ensures that the goals set for the program are appropriate and measurable. Also, that baseline data are collected at the beginning of the program, which can later be compared with the evaluation results. An ongoing, constructive evaluation increases the likelihood of credible results. Additionally, it proves cost efficient by taking advantage of the program infrastructure when planning evaluation activities (Yoong et al., 2013).

Ideally, evaluation plans are based on theory. *Program theory* should guide how evaluation questions are formulated and prioritized, how the research is designed and how the findings are interpreted. The literature lacks a consensus about the description and setup of program theory. Instead, there are various ways that have various names (Rossi et al., 2019). One way is not necessarily better than the other, and to some extent, it will be a question of taste which way and name to choose. In this paper, we present a simple scheme from Rossi et al. (2019), as it provides a good overview of what program theory should include. The scheme consists of three interrelated components: (1) the program's organizational plan, (2) the program's service utilization plan and (3) the program's theory of change (program impact theory).⁴

The *program's organizational plan* is developed by program managers and outlines the resources (human, financial and physical) as well as administrative and general organizational factors needed for the program. The plan comprises the functions and activities a program is supposed to perform given resources as well as the factors and preconditions needed for the organization to be able to offer the program (e.g. fundraising, acquisition and maintenance of facilities or governmental contacts). In general, the plan can be understood as a set of propositions in the following sense (Rossi et al., 2019, p. 67): "*If the program has such and such resources, facilities, personnel and so on, if it is organized and administered in such and such a manner, and if it engages in such and such activities and functions, then a viable organization will result that can operate the intended service delivery system.*"

The *program's service utilization plan* describes how the organization expects to reach its target population and also which service contacts are needed along the program. The plan "*pulls into focus the critical assumptions about how and why the intended recipients of service will actually become engaged with the program and follow through to the point of receiving sufficient services to initiate the change process represented in the program impact theory*" (Rossi et al., 2019, p. 68).

The *program's theory of change* describes the underlying logic model of the program, i.e. the causal theory of how the program is supposed to lead to the intended outcomes. It displays the sequence of causes, i.e. specific activities of the program, and effects that these activities are expected to produce, i.e. the social benefits. In most cases, a program can only indirectly impact the social conditions of interest. It does so by influencing crucial but controllable key aspects of the given situation, which in turn contribute to improvements of the social conditions. The theory of change goes through each step of the chain of events, revealing the assumptions about the change process and the respective dependencies to bring about the desired social benefits (Rossi et al., 2019). Having a clear understanding of a program's theory

⁴ Often the evaluation literature does not distinguish between the three components and discusses them together as the theory of change. Given their interrelations, this is a legitimate and practical way to proceed. However, we find it useful to, at least once, look at the components separately in order to understand the power of program theory and the many assumptions a program, and therefore also its evaluation, relies on.

of change is especially important for an impact evaluation, as its goal is to measure causal effects. The hands-on construction and relevance of a program's theory of change is discussed in Mauser (2025).

If a program is considered ineffective, then the theory of change, together with the organizational plan and the service utilization plan, can potentially explain why the program failed. For instance, by shedding light on whether specific activities were not implemented with the required quality or whether resources were allocated inadequately (Berry and Sloper, 2016). To this end, an exhaustive program theory needs to reveal all critical assumptions and expectations that underlie the program's design. Formulating program theory is not a one-shot job. The plausibility and logic of the proposed program theory need to be reviewed carefully as part of a continuous process, ideally throughout a program's lifetime (Rossi et al., 2019). *"A program whose design is weak or faulty has little prospect for success, even if it adequately implements that design. Thus, if the program theory is not sound, there may be little reason to assess other evaluation issues such as impact or efficiency"* (Rossi et al., 2019, p. 87). Hence, a poorly defined or flawed program theory makes program evaluation meaningless in the first place. For this reason, program theory itself needs to be assessed too.

One approach to determining the quality of program theory is an *evaluability assessment*. A decision-making tool (Craig and Campbell, 2015), it is commonly used to assess whether a program is ready for evaluation and to support evaluation planning as well as program development. At an earlier stage, an evaluability assessment can help identify promising interventions for evaluation (Lam and Skinner, 2021). According to Rossi et al. (2019), evaluability assessment consists of three main parts: (1) describing the program model with a focus on the definition of objectives and goals, (2) assessing to which degree the model is both well-defined and evaluable and (3) identifying stakeholders' interests in evaluation and usage of evaluation findings. Importantly, evaluability assessments should not be confused with the evaluation per se. Although insights from an evaluability assessment can serve formative purposes (program improvement), they do not serve summative purposes (program accountability) (Lam and Skinner, 2021).

Financial education programs can differ in various ways. Seldom are they entirely built on unique assumptions about how the program will function and how the intended changes are supposed to happen. As a first step of assessing evaluability, literature on social sciences and human services can provide evidence and insights from practical experiences, which can help assess the meaningfulness of the proposed program theory (Rossi et al., 2019). Besides a detailed desk review on information related to the history, design and characteristics of the program (i.e. *secondary data collection*)⁵, Yoong et al. (2013) advise to interview key stakeholders and conduct site visits and observations of program activities. Interviews do not only yield feedback (especially from the program staff) on the program and its operations but also inform about stakeholders' needs and their commitment to evaluating the program, as well as their potential contribution to evaluation (e.g. time, money, further information). Observing programs in action, additionally, helps check whether the resources, structures and processes necessary to conduct an evaluation are in place.

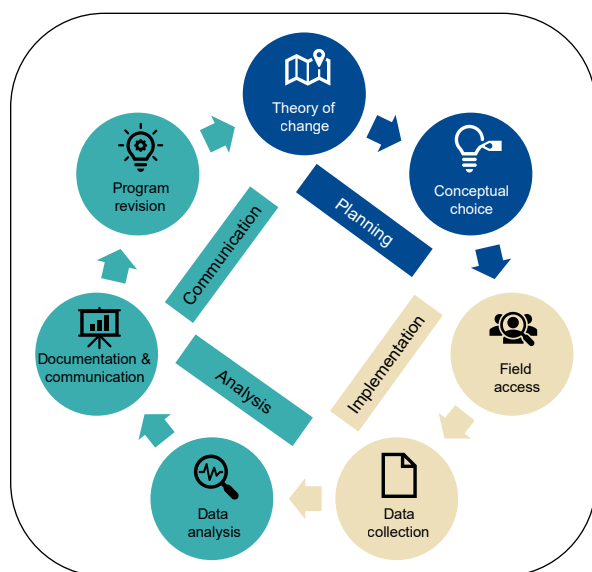
⁵ Secondary data collection refers to acquiring data that was not originally gathered for the evaluation at hand. These data come from secondary sources (Rossi et al., 2019).

While conceptualizations of evaluability assessments and their implementation vary (Davies, 2013), the fundamental premise is that they provide a systematic way to judge the desirability and feasibility of evaluation (Yoong et al., 2013).⁶ After all, the “results of such assessment should be reasonable consensus on a shortlist of potential evaluation questions and a preliminary understanding of the possible scope, methods and design of the evaluation to be deployed” (Yoong et al., 2013, p. 30). If the evaluability assessment concludes that an evaluation should be conducted, the next steps of the evaluation cycle can start.

Figure 1

3 The cycle of evaluation

Evaluation cycle



Source: Theresa Lorenz (OeNB).

The *evaluation cycle* describes the entire evaluation process and consists of four phases: (1) planning, (2) implementation, (3) analysis and (4) communication. Figure 1 gives an overview of these phases and the corresponding steps, which we will discuss in more detail below.

3.1 Planning

First, an evaluability assessment tells us whether to conduct an evaluation. Second, mapping out program theory, most importantly the theory of change, is already the first step of an evaluation’s planning phase. A program’s theory of change describes the logic model how certain inputs and activities generating certain outputs are supposed to lead to specific outcomes. It thus informs the decision on an appropriate evaluation concept. Table 1 summarizes the main (measurable) components of a theory of change and gives some examples.

Table 1

Theory of change – main components

Theory of change	Inputs	Activities	Outputs	Outcomes
Description	Resources necessary for implementing and operating the program	Specific actions and services that constitute the program	Direct and tangible products of the program	Effects on recipients which can be attributed to the program (short-, medium- and long-term)
Example	<ul style="list-style-type: none"> • Funding • Staff • Materials 	<ul style="list-style-type: none"> • Workshops • E-learning • Games 	<ul style="list-style-type: none"> • Number of participants • Hours of service • Number of awarded certificates 	<ul style="list-style-type: none"> • Changes in financial knowledge • attitudes • behaviors

Source: Author’s compilation based on Berry and Sloper (2016).

Looking at the examples given in table 1, we see that *inputs*, *activities* and *outputs* are measurable and quantifiable, e.g. the amount of funding, the number of e-learning modules a program includes or the number of participants who attended an offered workshop. To conduct program evaluation, however, it is essential to also identify measurable *outcomes*, which need to be achievable in a reasonable way within a

⁶ Discussing evaluability assessments in details is beyond the scope of this paper. For readers interested in the topic, we recommend having a look at Davies (2013), Lam and Skinner (2021) as well as Craig and Campbell (2015).

specified period and which require specific resources and activities (Berry and Sloper, 2016). As mentioned in section 1, intended program effects can be short-term, medium-term or long-term.

Often, short-term effects must be established first to create a foundation for more lasting effects. In this sense, outcome measures follow a hierarchical order (Berry and Sloper, 2016): (1) The lowest level of measurable outcomes is the recipients' satisfaction with the provided program. This is not only an important measure in process evaluation (box 2) but also essential for a program to have an impact. (2) Only if the recipients are satisfied with the program can the next outcome level be addressed, namely *learning to change* (short-term). This level includes changes in financial knowledge, attitudes and skills and recipients' aspiration toward desirable behaviors. (3) The next level is *taking action to change* (medium-term), which includes desirable changes in recipients' financial behavior and practices, e.g. establishing good routines. (4) The highest level of outcomes are *results of change* (long-term). From the individual perspective, the ultimate result is personal financial well-being, which, in turn, positively impacts on social, economic, societal and environmental aspects.

Based on the theory of change, evaluation questions can be formulated, laying out the evaluation priorities and goals. In box 1 and 2, we presented such evaluation questions, assigning them to formative vs. summative purposes and classifying them as part of process or impact evaluation. Having identified the components and respective measures of the theory of change, we can now assign the evaluation questions accordingly (box 3).

Box 3

Evaluation questions addressing components of the theory of change

Inputs

- Are the materials adequate and relevant for the recipients' needs and interests? (formative, process evaluation)
- Is the recipient-educator ratio appropriate for the workshop? (formative, process evaluation)

Activities

- Can the entire content be delivered to the recipients as intended? (formative, process evaluation)
- Are the program's materials used in an engaging manner? (formative, process evaluation)

Outputs

- Do the targeted recipients use the offered program? (formative, process evaluation)
- Does a sufficient number of the people in the program complete it? (formative, process evaluation)

Outcomes

- Are recipients satisfied with the provided program? (summative, process evaluation)
- Are there observable tendencies as to the desired outcomes (e.g. improved knowledge, greater awareness of financial risks, developing a budget plan)? (summative, impact evaluation)
- Does the program alleviate the problem it is supposed to address? If yes, to what extent? (summative, impact evaluation)

Importantly, evaluation questions should not be confused with survey questions (survey items). While the former outline the evaluation interest in a general sense, the latter are specific items used to examine changes in recipients (Berry and Sloper, 2016). For example, to help answer the evaluation question "Are

recipients satisfied with the provided program?”, a corresponding survey question addressed to the recipient could be “How satisfied were you with the content of the workshop (response options on a Likert scale from “very satisfied” to “not at all satisfied”)?”

“Evaluations should be issues-led not methods-led. And having determined the evaluation questions, the best available method should then be used to answer them” (White, 2010, p. 162). This brings us to the next step of planning, namely the conceptual choices of the evaluation design and method. Table 2 provides an overview of the key decisions to be made, addressing the following questions: (1) Which type of evaluation (see section 1) should be conducted and (2) how (evaluation design), (3) what type of data should be collected with which method and (4) who should the data be collected from (study sample)?

“A good evaluation design addresses the evaluation questions, is appropriate for the evaluation context (e.g., time, resources), and provides sufficient and critical data” (Berry and Sloper, 2016, p. 16). We differentiate between three types of evaluation design (Berry and Sloper, 2016): (1) *Non-experimental designs* are explorative (descriptive or correlational), looking at relationships among program implementation, participation and outcomes, for instance, using post-test only or examining pre-post changes. Non-experimental designs do not measure causal effects. This means that observed results cannot be argued to have been caused by the program (directionality), nor can they be directly attributed to it. (2) *Experimental designs* allow for measuring attribution and program impact, i.e. causality, by comparing outcomes between treatment and control groups. The treatment group consists of program recipients, i.e. individuals exposed to the program. The control group comprises individuals similar to those in the treatment group; yet, they do not participate in the program. Randomly assigning individuals from the study sample to these groups ensures that they have similar characteristics (except for program participation). This allows for causal conclusions based on observed differences between the groups. (3) *Quasi-experimental designs* also compare outcomes between program recipients and nonrecipients; however, random assignment is not possible. Instead, alternative methods are used to create comparable groups. The advantage of random assignment is that it eliminates threats to *internal validity* (establishing a causal relationship), whereas alternative methods can only reduce these threats.⁷

Table 2

Key decisions in the conceptual phase

Objective of evaluation	Program improvement	Program accountability
Evaluation type	Process evaluation <ul style="list-style-type: none"> Assess recipients' needs, adequacy of resources, quality of provided service Identify strengths and challenges of program implementation 	Impact evaluation <ul style="list-style-type: none"> Assess program's effect on recipients Identify what works for whom
Evaluation design	<ul style="list-style-type: none"> Non-experimental 	<ul style="list-style-type: none"> Experimental Quasi-experimental
Evaluation method	<ul style="list-style-type: none"> Needs assessment (quan/qual¹) Observations (qual/mix¹) Focus groups (qual) Surveys (quan/mix) Document review (qual/mix) Dosage/attendance (quan) 	<ul style="list-style-type: none"> Surveys (quan/mix) Tests or assessments (quan) Focus groups/interviews (qual)
Study sample	<ul style="list-style-type: none"> Program recipients Program staff/instructors 	<ul style="list-style-type: none"> Program recipients and a comparison group of non-recipients (control group)

Source: Author's compilation based on Berry and Sloper (2016).

¹ quan = quantitative method; qual = qualitative method; mix = mixed method.

⁷ Experimental and quasi-experimental designs will be discussed in an upcoming paper of the *OeNB Financial Literacy Evaluation Series*.

Looking at evaluation designs, we need to consider one more question (Berry and Sloper, 2016): When and how often should data be collected? Data can either be collected at one point in time (*cross-sectional design*), or data are collected at least at two points in time (*longitudinal design*). Cross-sectional designs have the disadvantage that they cannot observe changes over time, but they are commonly used to collect data after the program (post-test only) to examine differences between recipients and nonrecipients. Longitudinal designs are more challenging (e.g. time-consuming, costly, program dropouts) but have the advantage that they can explore changes in outcomes over time by comparing outcomes before the program (pre-test) to outcomes after the program (post-test).

Evaluation methods for data collection (mostly *primary data collection*⁸) depend on the nature of the data being gathered (Berry and Sloper, 2016; Rossi et al., 2019): *Quantitative data* involves numerical information, such as test scores, survey ratings and demographic data, which is useful for measuring program outcomes and ensuring accountability. Collecting quantitative data is expected to yield *valid* (accurate) and *reliable* (consistent) results. In contrast, *qualitative data* capture nonnumeric insights, such as written responses and personal narratives. Collecting qualitative data is more flexible and adaptive compared to quantitative data collection, representing experiences of those involved with the program and allowing for a deeper understanding of unintended effects and program dynamics.⁹ While both methods have their strengths and limitations, combining them in a *mixed-methods approach* can enhance the value and quality of evaluation. By integrating numerical data with recipient stories, evaluators can provide a more comprehensive picture of a program's impact (or implementation), balancing statistical evidence with human experiences.¹⁰ As with the evaluation design, the decision about which data to collect and which methods to employ depends on the evaluation questions as well as on the resources, including staff, funding and time, that are available for conducting the evaluation.

Depending on the evaluation design and method, the final step of the planning phase is to determine the study sample. The actual sampling process then takes place during the implementation phase, as part of accessing the field.¹¹ If all recipients of a program participate in data collection, a *full census* is achieved, which ensures complete representation. This approach is ideal when the total number of recipients is small and manageable or administrative data are available for outcome measures of interest. However, due to time or resource constraints, it is often not feasible to collect data from all recipients. In such cases, an evaluation sample must be selected (Berry and Sloper, 2016; Rossi et al., 2019). "*The primary goal of sample selection is to provide unbiased and sufficiently precise information on key study measures that adequately represent the target population for the evaluation. A sample is simply the subset of the units that make up the target population for an evaluation*" (Rossi et al., 2019, p. 273).

There are two main approaches to *sampling* (Berry and Sloper, 2016; Rossi et al., 2019): (1) *Probability samples* are selected randomly, which is to ensure that every member of the population has a known chance of inclusion. This method allows for unbiased, representative findings that can be generalized to the broader population. It also prevents selection bias, which can arise from human discretion in choosing recipients for the sample who portray the program in a particular light. However, probability sampling can be complex and resource-intensive. (2) When resources or time are limited, or when the evaluation is exploratory rather than aiming for broad generalizability, *non-probability sampling* (or *convenience sampling*)

⁸ Primary data refer to information directly gathered by the evaluators conducting the evaluation. However, also secondary data can be valuable for the evaluation, such as information from administrative databases (Rossi et al., 2019).

⁹ Qualitative research methods are discussed in Felbermayr (2024).

¹⁰ A practical guide for a mixed-methods approach is presented in Lorenz (2024).

¹¹ Details on survey methodology and sampling in quantitative research can be found in Zieser et al. (2025) and details on sampling in qualitative research in Felbermayr (2024).

is used. This involves selecting recipients based on convenience, specific characteristics or to ensure that diverse perspectives are included. While non-probability samples may not fully represent the population, evaluators aim to include key variations to capture a more comprehensive understanding of the program's impact.

To ensure a well-structured evaluation, careful sampling is essential (Rossi et al., 2019). Evaluators must determine who will provide data, how many recipients are feasible given the budget and timeline and how access to data will be secured. The selected sample should be at least reasonably representative of the target population, since evaluation results can be misled by sample populations which differ in many ways from the target population. Additionally, potentially missing data should be considered and addressed, such as accounting for non-responses by selecting a slightly larger initial sample (*oversampling*). By thoughtfully designing the sampling approach, evaluators can ensure that their findings are as meaningful and reliable as possible (Rossi et al., 2019).

3.2 Implementation

The implementation phase of an evaluation is often the most time-intensive, with *data collection* and management requiring significant planning and effort (Rossi et al., 2019). Even when secondary data sources are used, considerable time must be allocated for securing data access, ensuring ethical compliance, cleaning and integrating datasets and addressing data security concerns. Evaluations that involve primary data collection demand even more logistical coordination, particularly when *fieldwork* involves multiple sites, travel and engagement with various stakeholders (Rossi et al., 2019). Securing permissions, scheduling site visits, recruiting and training data collectors and processing collected data for analysis all contribute to the complexity of this phase. To keep the evaluation on track, a well-structured data collection and management plan is essential to ensuring that resources, staff and timelines align with the evaluation's objectives (Rossi et al., 2019).

“When designing evaluation tools and collecting data, special attention must be paid to protecting the rights of the participants. It is necessary to follow the human subjects governing rules and regulations to ensure that participants' privacy and freedom rights are not violated” (Berry and Sloper, 2016, p. 19). Lorenz and Felbermayr (2024) look at data privacy and research ethics in more detail. For now, we focus on the basic requirements for data collection and fieldwork, which apply to both quantitative and qualitative data collection.¹² Several critical factors must be considered when designing *evaluation tools* (data collection instruments and processes) to minimize errors, enhance accuracy and ensure meaningful insights: (1) Ensuring clarity and readability of survey questions, (2) addressing sensitive information, (3) minimizing the respondents' burden, (4) ensuring cultural appropriateness and relevance, (5) ensuring natural and unbiased measurement and (6) allowing for unexpected findings. Based on Berry and Sloper (2016) and MacDonald (2013), we summarize these factors in the following.

First, for data collection to yield valid and reliable results, survey questions must be clear, concise and easily understood by respondents. Ambiguous or complex wording can lead to misinterpretation, reducing the accuracy of responses. If a survey includes closed-ended questions (e.g. multiple-choice), all possible response options should be considered or an “other” option should be provided to avoid forcing respondents into inappropriate categories. Open-ended questions should be specific enough to guide meaningful responses while allowing respondents the flexibility to express their thoughts in their own words. For all survey questions, the reading level of the target audience must be considered to ensure that the language used matches the respondents' comprehension skills. This is especially important when working with children, individuals with low literacy levels or non-native speakers.

¹² For details on qualitative data collection, see Felbermayr (2024).

Second, when collecting sensitive data such as age, income or financial status, respondents may feel uncomfortable about providing precise information. To improve response rates and maintain comfort, questions should be designed in a way that protects privacy. One effective strategy is using response ranges rather than asking for exact figures. For example, instead of asking for a specific income, a question could present income brackets for respondents to select from. This approach fosters trust and encourages more honest responses.

Third, the process of data collection should be designed to minimize the burden on respondents. This includes considering the time and effort required to, for instance, complete surveys or participate in interviews. Overlapping evaluations within a single setting or population may result in respondents being surveyed multiple times, which may lead to survey fatigue. To avoid this, evaluators should explore opportunities for collaboration between organizations or stakeholders, sharing data collection efforts where possible to optimize resources while reducing redundancy for respondents.

Fourth, data collection methods should also be culturally appropriate to ensure that responses are accurate and meaningful. Survey questions should be sensitive to the values, traditions and norms of the population being studied. This requires engaging stakeholders in the design process to ensure that concepts are relevant and understandable in the given cultural context. Without these considerations, questions may be misunderstood or deemed irrelevant, which would compromise the integrity of the data.

Fifth, to maintain objectivity, survey questions should be written in a neutral manner without presupposing a particular outcome. For instance, instead of stating “increase in knowledge,” a more neutral phrasing would be “level of knowledge” to avoid implying a positive or negative result before data collection is complete. Not implying a preferred direction or value in the formulation of items ensures that the measurement remains unbiased, which allows for a fair and accurate assessment of the program’s impact.

Sixth, while evaluations often focus on predefined outcomes, it is also essential to build in flexibility to capture unanticipated effects of a program. Survey items should be structured in a way that allows for the documentation of unexpected trends or consequences, be it positive or negative. This adaptability enhances the depth of the evaluation by revealing insights beyond initial expectations.

Nightingale and Rossman (2015) point out another pitfall to be aware of when conducting field-based studies, namely the risk of collecting too much information. On the one hand, data collection is time- and resource-intensive, and on the other hand, an excessive amount of data can become overwhelming and disorienting in the subsequent data analysis phase.

To recall the most important principles of data collection, Gugerty and Karlan (2014; 2018) proposed the *CART* principles, where *CART* stands for credible, actionable, responsible and transportable data collection, summarized in box 4.¹³

¹³ For more details on how to develop effective survey items, we recommend reading chapter 3.3 in Yoong et al. (2013). There, the authors describe how to choose high-quality items according to the SMART principle, short for specific, measurable, attributable, realistic and targeted.

CART principles

Credible – Collect high-quality data that can be analyzed accurately

Data should be collected in a consistent way (reliability) and accurately reflect what one seeks to measure (validity). Subsequent credible analysis requires understanding of what can be measured (especially important when attempting estimates of impact).

Actionable – Collect data you can commit to use

Collect only the data that will be used. For every piece of data collected, ask: (1) Is there a specific action that will be taken based on the findings? (2) Are the necessary resources available to implement that action? (3) Does the commitment to take that action exist?

Responsible – Benefits of data collection should outweigh the costs

Data collection should be matched with the available systems and resources for collection, such that benefits exceed costs. Costs include direct costs of data collection but also opportunity costs, i.e. money and time that could have been used elsewhere. Data collection should not exceed the ability of data analysis. Responsible data collection requires transparent processes, protection of individuals' sensitive information and proper documentation.

Transportable – Collected data should generate knowledge

Valuable lessons generated from evaluations should help build more effective programs by applying what has been learned (why a program works) either to the program itself in the future or to other similar programs. Transportability requires transparency, i.e. sharing learnings with others.

3.3 Analysis

Before an evaluator can analyze the collected data, they must first be stored and documented. Thorough *data management* is needed to ensure that the collected data remain secure, accessible and well-documented for both current and future use. A well-structured data system should have sufficient storage capacity, secure backup processes and appropriate hardware and software to support data integrity (Yoong et al., 2013). Regular backups should be stored separately from primary data to prevent loss in case of system failures. When working with sensitive data, especially involving people, strict security measures are needed. Access to individually identifiable information should be restricted and require proper authorization and verification. Where possible, identifiers should be anonymized to protect personal information, and access logs should be maintained to track who interacts with the data (Yoong et al., 2013).

Comprehensive *data documentation* is also critical, as Yoong et al. (2013) note. Final datasets should be accompanied by detailed metadata, including variable definitions, sampling strategies, data collection methodologies and fieldwork details such as nonresponse rates. This ensures transparency, allowing users to accurately interpret, reproduce and apply the data for future evaluations. Proper documentation not only supports the integrity of the current evaluation but also serves as a valuable resource for future research and decision-making.

The *data analysis process* depends on whether the collected data are quantitative or qualitative. On a general level, analyzing and summarizing quantitative data involves calculating percentages and means (or other measures of central tendency). Presenting and visualizing data in a clear and understandable way

is also an important part of the data analysis process. This can be done, for instance, by using tables, graphs or bar and pie charts. Analyzing qualitative data typically involves examining responses to open-ended questions, success stories or respondents' observations. A simple way to summarize qualitative data is by identifying common themes and categorizing data accordingly, e.g. by grouping quotes thematically (Berry and Sloper, 2016). See Felbermayr (2024) for details on data analysis techniques for qualitative data and Lorenz et al. (2025a) for quantitative data.

Instead of going into further detail on data analysis processes, we close this section by discussing potential biases and noise that may arise during analysis and when interpreting the findings (Picciotto, 2022, Kahneman et al., 2021):

Bias refers to systematic deviations from objective judgment, often stemming from cognitive tendencies, ideological influences and structural dependencies. One major source of bias is *confirmation bias*, which occurs when evaluators prioritize evidence that supports pre-existing theories or stakeholder expectations while dismissing contradictory data. Another form of bias arises from *motivated reasoning* and *conflicts of interest*, particularly when evaluators are financially or institutionally dependent in their work. In such cases, they may be (indirectly) incentivized to align their findings with relevant stakeholders' expectations rather than conducting independent assessments. *Framing effects* also play a crucial role in shaping conclusions. Even the way survey questions are formulated can influence responses and interpretations, subtly steering judgments in a particular direction. Similarly, *status quo bias*, which is the tendency to favor existing practices and resist change, can prevent evaluators from considering innovative but potentially disruptive alternatives. As mentioned at the beginning of this paper, if political and public objectives compromise evaluation integrity, such biases can be indirectly reinforced.

While bias leads to systematic errors, *noise* represents inconsistent and unpredictable variations in judgment. This occurs when different evaluators, or even the same evaluator at different times, arrive at divergent conclusions based on the same data. One key source of noise is *bounded rationality*, which highlights the limitations in human cognitive capacity when processing complex information. Weighting multiple competing factors can be challenging and lead to inconsistent judgments that are influenced by external pressures, social constraints and information processing costs. Another common issue is *base rate neglect*, where evaluators fail to consider the underlying probability of an event when interpreting results. This can lead to overestimations of a program's success, particularly when positive outcomes are observed in a low-prevalence population. Similarly, the *illusion of validity* skews assessments when evaluators develop strong confidence in their judgments despite insufficient or contradictory evidence, such as when relying on confirmation bias. A further challenge is the influence of *availability cascades*, in which repeated exposure to certain interpretations or perspectives leads to their widespread acceptance, even if they are flawed. Positive-feedback mechanisms, e.g. through social dynamics and media reinforcement, can amplify these cascades.

Evaluation processes involve assessing a program's *merit* (doing things right), *worth* (doing the right things) and *value* (overall societal benefit) (Scriven, 1991). Each of these dimensions presents challenges that contribute to bias and noise. Merit assessments are often based on predefined standards, but when those standards are set by dominant power structures, they may reinforce existing inequities which can be (socially) counterproductive. Worth assessments address whether a program serves its intended beneficiaries but are inherently subjective, as different stakeholders may have conflicting priorities. Value assessments, which seek to integrate merit and worth into a broader judgment of social impact, are the most contentious, as they require trade-offs between competing goals, such as short-term efficiency versus long-term sustainability.

It is important to note that absolute objectivity in evaluation is unattainable. Evaluators must recognize the limitations of their knowledge and be aware that their judgments are inherently relative and tentative. By acknowledging the inevitable presence of bias and noise, evaluators can adopt strategies to mitigate negative effects, such as diversifying methodological approaches, fostering independence from vested interests and continuously reassessing assumptions. This, in turn, highlights the importance of reviewing program theory, for example, through evaluability assessments (see section 2). Ultimately, while evaluation cannot produce universal truths, it can provide meaningful insights that support informed decision-making for program improvement and accountability and, more broadly, serve the interests of society and the public good.

3.4 Communication

The final phase of the evaluation cycle focuses on documenting and communicating the evaluation findings. The ultimate goal is to use these findings to revise the program.

Like Rossi et al. (2019) point out, a *communication plan* is a crucial component of an evaluation and, strictly speaking, should already be considered during the planning phase. Its purpose is to determine how evaluation findings can effectively guide actions and attitudes in alignment with the evaluation's purpose while maintaining transparency about the process, data, findings and limitations, as well as ensuring evaluator independence. At its core, the communication plan outlines an agreement between the evaluator and key stakeholders (e.g. sponsors) on how evaluation findings will be released. A key consideration is determining who has the right to release the findings, as organizations commissioning evaluations often control the timing and extent of public disclosure. To safeguard evaluation integrity, evaluators may seek to retain intellectual property rights over the data, ensuring that results are communicated in full, regardless of whether the findings are favorable. Establishing clear agreements on information ownership and disclosure conditions during the evaluation's planning phase is essential to avoiding conflicts later, particularly if the results are sensitive or controversial. Without a well-defined communication plan, negotiations about releasing findings can become challenging, especially once the results are known.

Communication of evaluation findings is embedded in a complex environment, shaped by *multiple stakeholders*, who often have diverging, and at times conflicting, interests. These stakeholders, typically those with a direct and visible stake in a program, may hold differing views on both the legitimacy of the evaluation process and the potential consequences of its findings. Given these competing perspectives, evaluators must carefully navigate their relationships with stakeholders, ensuring that their role remains one of providing empirical evidence rather than making direct judgments. However, the *"distinction between making judgments and providing information on which judgments can be based is useful and clear in the abstract but often difficult to make in practice"* (Rossi et al., 2019, p. 294).

Evaluation involves *political and public interests*; therefore, evaluators must recognize that their contributions are just one of many factors influencing stakeholders' decision-making. Differing political or institutional perspectives can create tensions that evaluators must anticipate and plan for. However, *"the responsibility of the evaluator is not to take one of the many perspectives as the sole legitimate one but, rather, to be clear about the perspective from which a particular evaluation is being undertaken while giving recognition to the other perspectives"* (Rossi et al., 2019, p. 295). Stakeholder reactions to evaluation results can be challenging. Even well-grounded, rigorously conducted evaluations may be perceived as biased or arbitrary, especially if the findings contradict stakeholders' interests, which leads to resistance or controversy. Even those who commission an evaluation may reject or discredit the results, potentially undermining the evaluator's professional standing or future opportunities. This puts evaluators in a precarious position, particularly when sponsors or influential stakeholders withdraw support due to unfavorable results.

Another major difficulty arises from *communication barriers*, as evaluation terminology, rooted in social science disciplines, can be difficult for nonexpert audiences to grasp. Ensuring that findings are accessible and clearly communicated is crucial for fostering understanding and encouraging the practical application of results. The diverse backgrounds of stakeholders make this a persistent challenge (Rossi et al., 2019). “Documenting and communicating your results is the key to ensuring that interested audiences – including program participants and managers, members of your organization, funders and organizations that operate similar programs – can learn from what you found in your evaluation and apply the results” (Yoong et al., 2013, p. 214). Evaluation results are not only valuable to program stakeholders but also contribute to generating knowledge for academics in evaluation science, as described in section 1. However, “learning from knowledge” can be compromised if not all evaluation findings are documented and communicated. The tendency to publish only results that show large program effects, while neglecting those with weak, negative or no effects, is known as *publication bias* (Yoong et al., 2013). Evaluating financial education programs, particularly their impact, provides valuable insights for future program development. As Yoong et al. (2013) emphasize, it is essential that the evidence base includes all types of findings, not just those showing large effects, to present a balanced and accurate picture of actual program outcomes.

Based on a comprehensive communication plan, writing the report can start once the evaluation findings are available. One way of writing up the findings can be in the form of a *formal evaluation report*. Such a report is often required by program oversight decision-makers and program sponsors. Generally, it is important to approach the report strategically with a view to ensuring it effectively reaches and engages the intended audiences. The structure of a formal report should be clear and purposeful, guiding readers through the key findings, analysis and implications in a way that is both accessible and actionable. We summarize the six main parts of a formal report as described by Yoong et al. (2013) in the following.

First, although mostly written at the end of the report, the *summary* comes at the very beginning of the final report. It should provide an overview of the context, evaluation approach, key findings and recommendations for action. The summary should be brief and self-contained, ensuring that even those who do not go through the full report grasp the essential points.

Second, the *introduction* serves as a roadmap, giving readers an overview of what to expect in the report. It should provide an outline of the program and its purpose, the evaluation’s objectives, the evaluation stakeholders and their relationship to the evaluators and the methodological approach.

Third, the report should contain a detailed *program description*, which presents the program’s history, background and development, the program’s objectives and goals and also the activities offered and targeted recipients.

Fourth and following this section, the *evaluation design and methods* should be described. This section should clearly state the evaluation questions and the data collection methods used to address these questions. In addition, the analysis methods for each type of data collected should likewise be outlined.

Fifth, presenting *evaluation results* requires synthesizing data into a meaningful narrative. To make it easy for readers to follow, this section should align findings with the evaluation questions (or specific evaluation tasks). In cases where quantitative analysis plays a key role, the writing process may also involve making interpretative decisions to extract clear themes or messages that translate into actionable insights. To improve readability, incorporating visual elements (see section 3.3) is particularly effective. To ensure clarity, any visuals must be accompanied by well-written explanations. Having unclear visual representations or descriptions can prove counterproductive, as they may confuse readers and disrupt the flow of the report. Apart from clarity, sensitivity in presenting findings is critical, especially if the evaluation reveals shortcomings in program performance, staff effectiveness or operational efficiency. The results should be framed objectively and tactfully, with the authors avoiding language by which stakeholders might feel directly criticized. Policy audiences and funders might benefit from a highly

condensed version of the result section, which is sometimes referred to as “microcontent.” This might include a brief, bulleted summary alongside a compelling visual, designed to fit on a webpage or be used as a standalone document.

Sixth, the *conclusions and recommendations* section is particularly important for audiences expected to take action based on the report. Naturally, this section should build on the results. In other words, authors should ensure that the conclusions are directly supported by the preceding findings (vs. unfounded statements). Additionally, recommendations should not only address current issues but also point to future steps or improvements. Clearly articulating the connections between findings, conclusions and suggested actions supports the credibility and utility of the report.

Once the report is finalized, the final step is ensuring it reaches the right audiences, as underlined by Yoong et al. (2013). This involves selecting the most effective *dissemination channels*, which can include in-person presentations, print distribution and digital platforms. In today’s media-rich environment, leveraging multiple modes of communication can be demanding but enhances the report’s reach and impact. Overall, careful attention should be given to how and where the findings are shared to maximize engagement and influence decision-making.¹⁴

4 Summary and concluding remarks

Program evaluation plays a vital role in determining the effectiveness of financial education programs. This paper outlines the evaluation cycle, covering the key phases of planning, implementation, analysis and communication. It focuses on the methodological and strategic considerations necessary for a thorough and objective evaluation.

In the planning phase, defining a clear evaluation purpose and setting measurable objectives are essential. The evaluation design should align with the program’s stage of development, distinguishing between formative and summative evaluation goals. Formative evaluation focuses on assessing program implementation and refining processes, while summative evaluation examines program outcomes and impact. As broader categories, process and impact evaluation consider both implementation quality and program effectiveness, aligning with the needs of different decision-makers. Process evaluation is primarily used by program administrators to refine operations, whereas impact evaluation informs policymakers and funders about long-term program success.

During the implementation phase, careful data collection is crucial to ensure reliable and valid findings. The choice of evaluation methods, be it qualitative, quantitative or mixed methods, must be guided by the evaluation questions and available resources. Sampling strategies, ethical considerations and logistical challenges also play a significant role in shaping the evaluation process. A well-structured approach to data collection, aligned with stakeholder needs and practical constraints, enhances the credibility of findings.

The analysis phase involves synthesizing collected data to derive meaningful conclusions. Quantitative analysis may include statistical comparisons of program outcomes, while qualitative analysis provides deeper insights into participants’ experiences. Interpreting the findings is not without challenges. Biases, such as confirmation bias, motivated reasoning and framing effects, can influence how results are presented and understood. Additionally, noise in evaluation judgments can lead to inconsistencies. Recognizing these potential pitfalls and applying rigorous analytical frameworks helps ensure that findings accurately reflect program effectiveness.

¹⁴ A more elaborate description of the communication process is beyond the scope of this paper. Interested readers may refer to chapter 14 in Yoong et al. (2013) and the study by da Costa (2012) for further insights.

The final phase, communication, determines how evaluation findings are conveyed to stakeholders. A structured communication plan is crucial to ensure that results reach decision-makers, funders and other interested parties in a clear and actionable manner. A formal evaluation report summarizes the findings. To maintain objectivity and stakeholder engagement, careful attention must be paid to how results are framed, particularly with regard to challenges or unfavorable findings.

In conclusion, program evaluation is an ongoing, iterative process that enhances the effectiveness and accountability of financial education initiatives. How can decision-makers generate meaningful evidence to inform future program development? They can do so by systematically following the evaluation cycle – from planning and implementation to analysis and communication. Moreover, evaluations contribute to the broader research field, thus helping establish best practices and methods for assessing financial education programs.

References

- Berry, T. and M. Sloper. 2016.** NEFE Financial Education Evaluation Manual. National Endowment for Financial Education. <https://toolkit.nefe.org/evaluation-resources/evaluation-manual/section-1-introduction/introduction-nefes-financial-education-evaluation-manual>
- Chambers, R., D. Karlan, M. Ravallion and P. Rogers. 2009.** Designing impact evaluations: different perspectives. In: International Initiative for Impact Evaluation 3ie Working Paper Series.
- Craig, P. and M. Campbell. 2015.** Evaluability Assessment: a systematic approach to deciding whether and how to evaluate programmes and policies. What Works Scotland. <https://whatworksscotland.ac.uk/>
- da Costa, P. 2012.** Study on Communicating Evaluation Results. Prepared for the OECD Informal Network of DAC Development Communicators (DevCom Network).
- Davies, R. 2013.** Planning Evaluability Assessments, A Synthesis of the Literature with Recommendations. Cambridge: Department for International Development.
- Felbermayr, K. 2024.** Qualitative research evaluation – how to get from first ideas to a final paper. OeNB Financial Literacy Evaluation Series.
- Gugerty, M. K. and D. Karlan. 2014.** Measuring Impact Isn't for Everyone. In: Stanford Social Innovation Review.
- Gugerty, M. K. and D. Karlan. 2018.** Ten Reasons Not to Measure Impact – and What to Do Instead. In: Stanford Social Innovation Review. 41–47.
- Kahneman, D., O. Sibony and C. R. Sunstein. 2021.** Noise: A Flaw in Human Judgment. HarperCollins Publishers.
- Lam, S. and K. Skinner. 2021.** The Use of Evaluability Assessments in Improving Future Evaluations: A Scoping Review of 10 Years of Literature (2008–2018). In: American Journal of Evaluation 42(4). 523–540.
- Lorenz, T. 2024.** Mixed methods – a practical guide for the gold standard of evaluation research. OeNB Financial Literacy Evaluation Series.
- Lorenz, T. and K. Felbermayr. 2024.** Data privacy and research ethics in financial literacy evaluation research. OeNB Financial Literacy Evaluation Series.
- Lorenz, T., S. Anyfantaki and M. Zieser. 2025a.** Quantitative data analysis in financial literacy evaluation research – visualization and statistical inference. OeNB Financial Literacy Evaluation Series.
- Lorenz, T., S. Mauser and M. Zieser. 2025b.** Is financial knowledge enough? Reviewing impact assessments of financial education interventions. OeNB Financial Literacy Evaluation Series.
- MacDonald, G. 2013.** Criteria for Selection of High-Performing Indicators, A Checklist to Inform Monitoring and Evaluation. Western Michigan University – The Evaluation Center. <https://wmich.edu/evaluation/checklists>
- Mauser, S. 2025.** Theory of change – understanding the link between a program's design and its goals. OeNB Financial Literacy Evaluation Series.
- Nightingale, D. S. and S. B. Rossman. 2015.** Collecting data in the field. In: K. E. Newcomer, H. P. Hatry and J. S. Wholey (eds). Handbook of practical program evaluation. 4th edition. New Jersey: Jossey bass. 445–473.
- Picciotto, R. 2022.** The psychology of evaluation. In: Evaluation and Program Planning 94.
- Rossi, P. H., M. W. Lipsey and G. T. Henry. 2019.** Evaluation, A Systematic Approach. 8th edition. SAGE Publications, Inc.
- Scriven, M. 1991.** Evaluation Thesaurus. London: Sage Publications.
- White, H. 2010.** A Contribution to Current Debates in Impact Evaluation. In: Evaluation 16(2). 153–164.
- Yoong, J., K. Mihaly, S. Bauhoff, L. Rabinovich and A. Hung. 2013.** A toolkit for the evaluation of financial capability programs in low- and middle-income countries. Washington DC: International Bank for Reconstruction and Development / The World Bank.
- Zieser M., T. Lorenz and V. Voith. 2025.** Conducting representative surveys and impact assessments: an overview of survey methodology. OeNB Financial Literacy Evaluation Series.