HEQCO's Program Evaluation Manual

For Programs Related to Access and Retention

Helen Tewolde and Jess McKeown

Cite this publication in the following format:

Tewolde, H. & McKeown, J. (2018)*. HEQCO's Program Evaluation Manual – For Programs Related to Access and Retention.* Toronto: Higher Education Quality Council of Ontario.

Acknowledgement:

Introduction

The Higher Education Quality Council of Ontario (HEQCO) is an agency of the Government of Ontario that advances evidence-based research for the continued improvement of the postsecondary education (PSE) system in Ontario. To enhance the access, quality and accountability of Ontario's colleges and universities, HEQCO supports a variety of program-evaluation projects in the PSE sector to ensure that promising practices are recognized and advanced through program and policy development.

Program evaluation is a priority at HEQCO for many reasons. The Government of Ontario spends millions of dollars each year on programs and initiatives intended to support student access and retention in postsecondary education. We need to ask ourselves: What works? What is our target? What are we measuring? Without consistent evaluations of publicly funded programs, it is impossible to answer these questions.

The Canadian Evaluation Society defines program evaluation as the systematic assessment of the design, implementation or results of an initiative for the purposes of learning or decision-making.[1] Program evaluations can help an organization or institution understand: a) if, how and why a program works the way it does; b) whether a program is achieving the intended outcomes; c) how a program might be improved, and/or d) whether a program can or should be scaled to other contexts. With such a broad definition, it is easy to see why program evaluations will drastically differ across contexts.

This manual aims to provide a concise overview of practical considerations and a guided process for those less familiar with planning and implementing a program evaluation. The primary intended audience for this manual are proponents of HEQCO's Access and Retention Consortium (ARC), a community of practice for those evaluating programs that enhance student participation and success in PSE. This manual can also be used more generally by those who are interested in evaluating related programs across the higher education sector at large. The manual has been divided into the following sections:

A. Establishing the purpose and scope of an evaluation:
1. Preparing for an evaluation by developing research questions
2. Defining key variables
3. Establishing project scope and who should be involved in an evaluation
4. Selecting an evaluation approach
5. Choosing a study design

B. Performing an evaluation:
6. Collecting the data
7. Sampling the population
8. Managing, storing and analyzing the data
9. Reporting results

C. Additional factors to consider for evaluation:
10. Meeting ethical standards
11. Building a budget

Sections one through nine have been organized as one possible sequence for planning an evaluation, but these steps will rarely occur in a linear fashion. Evaluators will find that program evaluation is an extremely iterative

---

1 For more detail on this definition of program evaluation, see the Canadian Evaluation Society's underline website.

process. Steps taken throughout the planning process may be completed simultaneously, or even in a different order than has been presented here. Depending on one's experience with evaluation, and potential role within a larger evaluation team, some sections will be more applicable than others. Whether evaluators are new to the field or have years of evaluation experience, they may find themselves working backward, revisiting earlier steps to clarify work happening later on, and changing their initial plan for execution. Planning extra time for the iterative nature of program evaluation is essential, especially from a project management perspective.

Before diving into the content of the manual, we note a few important caveats. First, regardless of one's preferred approach to program evaluation, the research questions are the most central and essential component of evaluation. The research questions, sometimes also referred to as evaluation questions, will be used to guide the entire evaluation process, from planning to implementation to reporting. This raises an important point about language. The semantics around program evaluation can differ from evaluator to evaluator, especially across different disciplines and sectors (e.g., health, psychology, education, etc.). We want to acknowledge that there is often more than one way of saying the same thing, and that in some cases, we have grouped similar concepts together to ease understanding for beginners. For simplicity's sake, we have sought to align the language used throughout this manual with what is most widely used in education. Our third caveat is that we will use the term "program" to refer to any program, intervention or strategy being evaluated. And lastly, given the complexity of program evaluation as a field of study and practice, it should be noted that this manual neither intends to, nor could ever achieve the goal of being comprehensive. Although this beginner's manual intends to simplify evaluation as much as possible, the inherent messiness of evaluation must still be acknowledged. There is never only one way of looking at things, and evaluation will never be as simple as it might appear to be on paper. We can only hope that this manual will be a useful resource for those new to evaluation, and we will direct to additional resources for certain topics when applicable.

A Establishing the Purpose and Scope of an Evaluation

Part 1: Preparing for an Evaluation by Developing Research Questions

Foundational to every program evaluation are the research questions. Research questions are essential for framing an evaluation because they are purpose driven, and will therefore inform all future choices that will be made throughout the project. Those involved in the evaluation will compare the many advantages and disadvantages for all associated research activities (e.g., study design, data collection, sampling techniques, data analysis, etc.), ultimately relying on the research questions to justify their choices. The number, focus and specificity of evaluation questions will depend on the experience of the evaluation team (Part 3). Having a few high-level research questions may be sufficient for expert evaluators, whereas those newer to evaluation may want to develop more specific questions. The following lines of questioning represent broad examples for research questions:

> **Research Questions**
> Foundational questions that articulate why a program is being evaluated and guide research activities
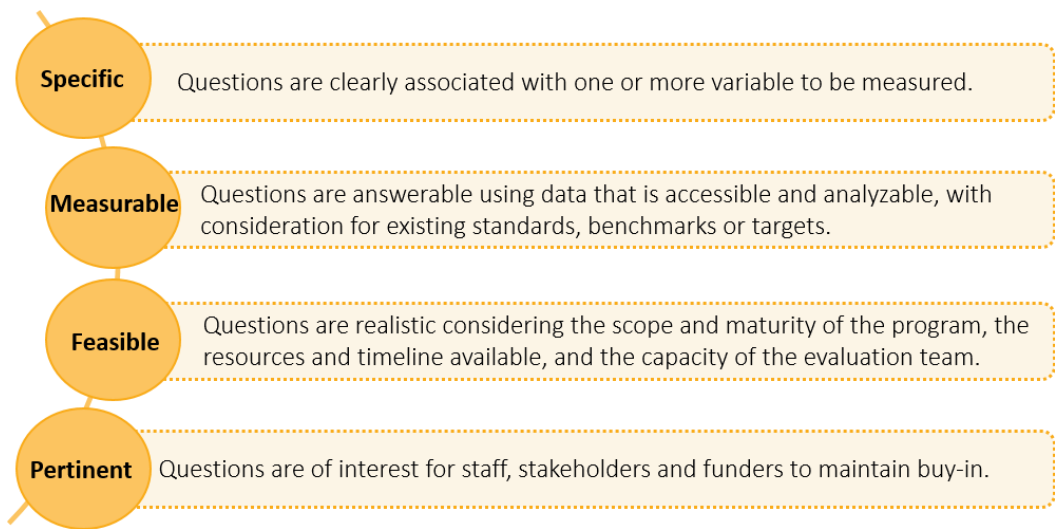
- ✓ Is the problem the program is trying to address well established and is the program consistent with the organization's mandate/mission?
- ✓ To what extent does the program align with the research and practical knowledge base about the problem and how it might be addressed?
- ✓ Has the program been implemented according to plan and to the intended target group?
- ✓ How or why does the program work the way it does?
- ✓ Is the program achieving the intended outcomes/goals (be they short-, medium- or long-term)?
- ✓ Are there any unintended or undesirable consequences of the program?
- ✓ How might the program be improved in future?
- ✓ To what extent does the program offer value for money when compared to other programs with similar aims?
- ✓ Can or should the program be transferred to other contexts and/or scaled up?
- ✓ How might similar programs in the sector learn from evaluation results?

It is very rare that an evaluation could address all areas of potential interest, so evaluators should think carefully about the most important aspects of a program to evaluate, possibly consulting relevant stakeholders when applicable (Part 3).[2] For evaluations that have multiple research questions, we recommend that questions are distinct yet complementary, and that the evaluator(s) map out the relationships and chronological progression through each. Ideally, well-articulated research questions will meet the criteria in Figure 1.[3] These criteria will be revisited throughout this document.
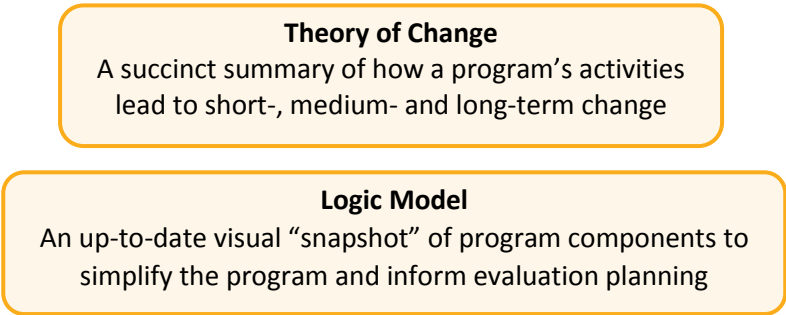
---

2 For more information on choosing research questions, see the Community Toolbox website here.
3 For more information on criteria for well-written research questions, see Wingate & Schroeter's (2007) resource here.

Figure 1: Criteria for Well-written Research Question(s).

Specific — Questions are clearly associated with one or more variable to be measured.

Measurable — Questions are answerable using data that is accessible and analyzable, with consideration for existing standards, benchmarks or targets.

Feasible — Questions are realistic considering the scope and maturity of the program, the resources and timeline available, and the capacity of the evaluation team.

Pertinent — Questions are of interest for staff, stakeholders and funders to maintain buy-in.

When and how research questions are developed will depend on the nature of the project. For instance, questions may have been predetermined prior to initiation of the evaluation, developed organically through conversations with program staff or constructed through serious deliberation with a group of senior directors. To support the development of research questions, it may be helpful to develop a theory of change and a logic model[4] for the program.

Theory of Change
A succinct summary of how a program's activities lead to short-, medium- and long-term change

Logic Model
An up-to-date visual "snapshot" of program components to simplify the program and inform evaluation planning

These tools can provide clarity for staff, stakeholders, evaluators and funders on a number of factors including the assumptions upon which the program is grounded, the context within which the program is implemented, and the core aspects for program function. Logic models are especially helpful for establishing, distinguishing and orienting relationships between program components, activities, outputs and outcomes for all parties involved in an evaluation (Figure 2). Differentiating these program components and noting other external factors that affect a program's implementation are critical for thinking about what will be measured during an evaluation. It should be noted that there are numerous variations of logic models that can be used to visualize a program in other unique and innovative ways. Reaching agreement on the conceptual understanding of one's program is more important than the specific design that is used. Examples of traditional logic models for ARC-related programs are depicted in Figures 3 and 4.

---

4 For more information on strategies and approaches to developing a theory of change and logic models, see BetterEvaluation's website here and/or the Government of Canada's website here.

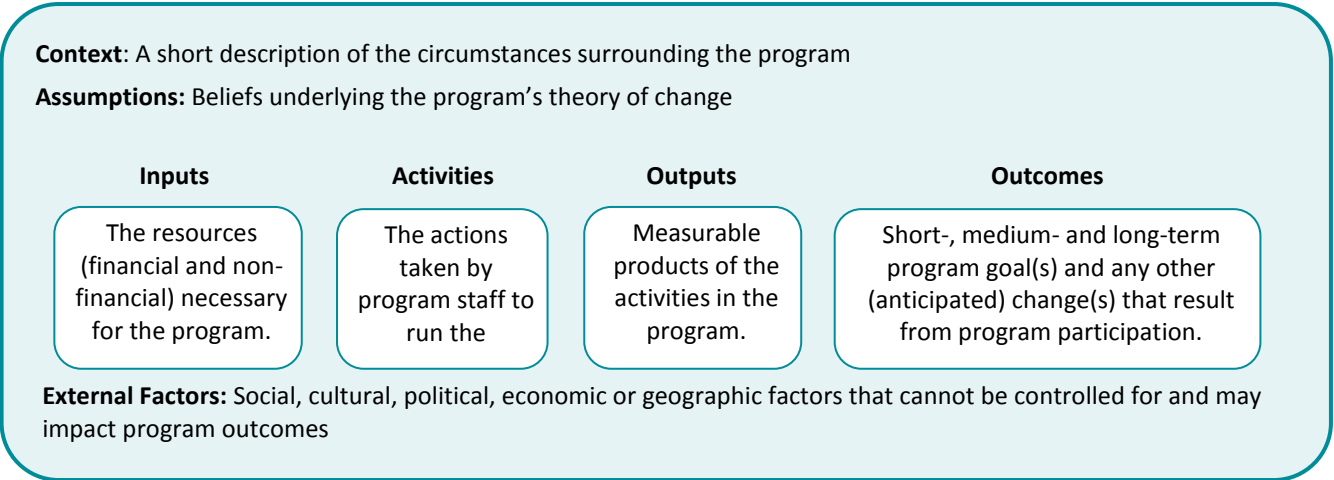**Figure 2: Traditional Logic Model Template Featuring Descriptions of Each Element**

Context: A short description of the circumstances surrounding the program

Assumptions: Beliefs underlying the program's theory of change

Inputs	Activities	Outputs	Outcomes
The resources (financial and non-financial) necessary for the program.	The actions taken by program staff to run the	Measurable products of the activities in the program.	Short-, medium- and long-term program goal(s) and any other (anticipated) change(s) that result from program participation.

External Factors: Social, cultural, political, economic or geographic factors that cannot be controlled for and may impact program outcomes

Figure 3: A Logic Model for a Program that Fully Subsidizes Low-income Students' PSE Application Fees

Context: Students from low-income households are less likely to apply for PSE than their wealthier peers. Completing PSE can enable social mobility and provide students with opportunities they would not otherwise have.

Assumptions: Low-income students are not applying to PSE because the application fees are prohibitively high; if PSE applications were completely subsidized, more students would apply, attend and complete PSE as a result.

Inputs	Activities	Outputs	Outcomes		
			Short-term	*Medium-term*	*Long-term*
Program staff	Student outreach and marketing	50 low-income students' PSE applications subsidized each year	More low-income students **apply to** PSE	More low-income students **attend** PSE	More low-income students **complete** PSE
Subsidy funding					
Participating students	Subsidy facilitation				

External Factors: Students' desire to apply to and attend PSE; students' qualifications; students' financial ability to attend PSE

Figure 4: A Logic Model for UTSC's ARC Project to Evaluate the Get Started Orientation Program

Context: Students and their caregivers can benefit from support when transitioning to their first year of PSE. Providing administrative guidance and other information via a free one-day orientation program can assist with the transition, and in turn increase students' retention and academic success.

Assumptions: Academic and social integration, and a sense of belonging on campus impact student academic achievement and retention.

Inputs	**Activities**	**Outputs**	**Outcomes**		
			Short-term	*Medium-term*	*Long-term*
Program staff and volunteer peer coaches	Administrative guidance on: • Course selection and scheduling • Student cards	Students obtain student card and course schedule	Students enroll in the appropriate courses	Students successfully transition to PSE and become integrated within the institution	Students have increased retention and academic success between year 1 and year 2
Program funding			Students learn about transitioning to PSE, institutional expectations and academic skills		
Participating students	Presentations on important topics: • Transitioning to PSE • Academic policies • Skills for academic success • Role of caregiver • Finances	Students and caregivers attend the *Get Started* presentation		Caregivers provide adequate support for students	
Participating caregivers			Caregivers learn about the various ways students may require support		

External Factors: Engagement with materials offered at *Get Started*; student motivation; degree of parental involvement; students' capacity; personal lives of students.

Once a program has been mapped out using a logic model, the visual can be used as a guide to determine what components of the program are of greatest interest for the evaluation. This may be a good time to review other evaluations that may have been conducted on similar programs, either formally by reviewing the literature or informally through conversations with other evaluators and program staff.

Now is a good time to reflect on how evaluation results will be used and the primary intended audience for the evaluation results (discussed further in Part 3). If the evaluation team cannot reach agreement about what portions of a program are to be evaluated, and how to structure research questions accordingly, there are several strategies that can be used to prompt discussion. For instance, performing a *SWOT analysis* is one common approach for narrowing in on the components of interest in an evaluation.[5]

> **SWOT Analysis**
> A reflection on the **S**trengths, **W**eaknesses, **O**pportunities and **T**hreats of one's program, intended to support the development of research questions

In the upcoming sections, we will discuss how research questions are to be used as the basis for many other choices to be made about the evaluation design.

---

5 For those interested in an example of a SWOT analysis, please see BetterEvaluation's website.

Part 2: Defining Key Variables

Before moving forward with the evaluation, there should be thoughtful consideration given to how each research question will be answered. To do this, key variables should be selected and defined.

> **Variables**
> Tangible indicators of the concept or outcome being measured that can be classified as either direct or indirect

As mentioned in Figure 1, research questions and associated variables can be vetted using the following four criteria: specific, measurable, feasible and pertinent. The process of identifying variables is effectively testing whether the research questions are *specific* and *measurable* (Figure 1). For some research questions, there may be a key variable that is easy to measure, whereas in other cases, more complex questions may necessitate a combination of multiple and/or proxy (i.e., indirect) variables. For instance, subjective concepts such as students' satisfaction with their first-year experience would be harder to measure than objective concepts, such as an institutions' retention rate between first and second year. The use of multiple and/or proxy variables is quite common in program evaluation, as singular variables are often hard to come by. That being said, evaluators are cautioned against using more than four proxy variables per research question, as this may imply that the concept/outcome being assessed should be better defined.

Specific

Measurable

Feasible

Pertinent
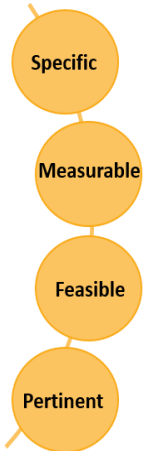
Case Consideration: Establishing variables to answer the research question

Let's say an evaluation of the application subsidy program from Figure 3 seeks to assess the short-term program outcome of increasing low-income students' applications to PSE. The use of PSE application data as a variable is straightforward and easy to retrieve, especially when such data is centralized. However, establishing students' low-income status represents a more difficult variable to measure, as many students may be unaware of their household income. To indirectly measure this same information, evaluators might create a proxy, whereby they collect students' postal codes and cross match these with census data to determine an average household income by area code (see Table 1).

When using multiple and/or proxy variables, it can be especially helpful to map out which variables will serve to answer each research question. This process might simultaneously prompt thinking about one's population of interest as well as data collection methods and sources, but these topics will be discussed further in Parts 6 and 7. Table 1 uses a potential research question from the evaluation described in Figure 3 to demonstrate.

Table 1: Identifying Variables for Research Question(s)[6]

Research Question	Variable(s)	Data Collection Method & Source
Did the application subsidy program contribute to an increase in the number of low-income students *applying* to PSE?	Number of PSE applications	Analysis of Ontario's Colleges'/Universities' Application Databases (OCAS & OUAC)
	Income quintile via: (a) Postal code (b) Census data	Document analysis of: (a) PSE admin data (b) Statistics Canada Census data

Once the evaluation team is satisfied that research questions and associated variables are *specific* and *measurable*, the next consideration is regarding the *feasibility* of collecting data on each variable (Figure 1). That is, researchers should consider the scope and maturity of the program, the resources and timeline available, and the capacity of the evaluation team. Considering the scope and maturity of the program is very important: Generally, programs in the early stages of implementation will be more suited towards process evaluations, whereas mature programs could be evaluated for either the process or the outcomes (Part 4).

Case Consideration: Examining the feasibility of research questions and associated variables

Suppose HEQCO is tasked with evaluating province-wide outcomes for the application subsidy program described in Figure 3. If HEQCO were to compare the feasibility of examining the short-, medium- and long-term program outcomes (i.e., student applications to PSE, student attendance to PSE and student completion rates from PSE, respectively) there are multiple factors to consider. Logistically, collecting PSE application data from a centralized database (OUAC/OCAS) might appear to be considerably less resource-intensive than collecting attendance/graduation data from each institution individually. However, feasibility also depends on who has access to what data; stakeholder relationships and associated data sharing agreements cannot be overlooked. In this case, HEQCO would expect that institutional researchers/administrators may have easier access to institutional attendance and graduation data, whereas OUAC or OCAS employees may have easier access to application data. Evidently, the selection of feasible research question(s) and associated variables would be dependent not only on the resources available, but also the relationships with those who have access to the required data.

The last consideration for research questions and associated variables is that they are *pertinent*; i.e., of interest and importance to the intended audience. Once evaluators have an idea of what data is available to them, they can consider what questions can be answered with this data. *Quantitative* (i.e., numerical) data is often useful to understand what is happening in a program, whereas *qualitative* (i.e., narrative/anecdotal) data might be considered more useful for explaining the extent to which this is happening, or why a program functions the way it does. Selecting variables can be quite an iterative process whereby evaluators may want to consult their audience about their data preferences, as their perceptions of the data can significantly influence the perceived credibility of findings. While both types of data — quantitative and qualitative — can be advantageous in different contexts, the best approach for adequately answering all research questions is usually to draw on both types of information. The next section will go into more detail about stakeholder engagement, during which time the pertinence of research questions and associated variables can be considered further.

---

6 For more information on identifying variables for research questions, see the Government of Canada's website here.

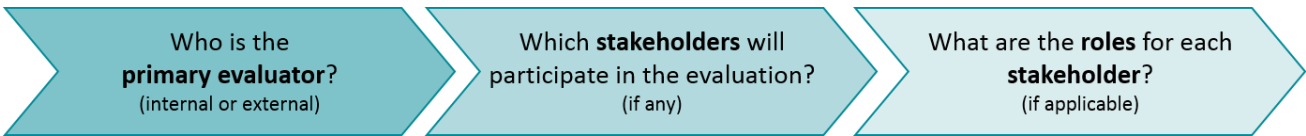# Part 3: Establishing Project Scope and Who Should be Involved in an Evaluation

Once research questions and associated variables have been established, it is a good time to determine who will be involved in the evaluation project. Stakeholders may be defined differently across contexts, but for the purposes of this manual, they will include all those who are affected by or interested in a program, whether these individuals are internal or external to that program. Gauging stakeholder interest in an evaluation, and defining their roles in the project is important so that selected research question(s) can be appraised as to whether they are *pertinent*. This will be especially important for institutional partnerships involving data sharing agreements, under which data and resources are combined to answer a variety of questions of interest for different groups. In any case, evaluators may find that initial research questions will be significantly refined once these are reviewed by applicable stakeholders, demonstrating the iterative nature of evaluation planning. Before diving into stakeholder considerations, it is recommended that the scope of the evaluation project is established. This might be done by answering the following questions:

- ✓ Who will use the evaluation results? And for what purpose will the results be used?
- ✓ What is the approximate time frame and budget? What other applicable resources are available?
- ✓ What capacity does the evaluation team have or require with respect to conducting the evaluation?
- ✓ Who needs to be engaged in the evaluation, and in what way, in order for the results to be useful?

Establishing the scope and defining the capacity required for an evaluation defines what is actually achievable within a specific time frame and budget (see Part 11 for more information on budget). Ultimately, it is important to avoid being sidetracked by idealism or an overestimate of what can be done, as these are common pitfalls in evaluation projects.

When considering who needs to be involved in an evaluation, the main consideration is the purpose of the evaluation. For programs that intend to consistently monitor post-evaluation program outcomes, developing early buy-in and capacity of program staff is essential. If the purpose is to have results that are meaningful and usable to a specific stakeholder, then including them from the beginning may help with that goal. If the purpose of the evaluation is more academic or does not require stakeholder buy-in, then it may be more useful to save the time and effort on stakeholder engagement for other elements of the evaluation.

Determining who should be involved in an evaluation can be done by answering the following questions:

Who is the **primary evaluator**? (internal or external)	Which **stakeholders** will participate in the evaluation? (if any)	What are the **roles** for each **stakeholder**? (if applicable)

To identify a primary evaluator, the first decision to be made is whether this individual will be internal or external to the program being evaluated. Internal evaluators are selected from within an organization/institution usually due to one or more of the following reasons: (1) internal evaluation capacity exists; (2) the program is straightforward; (3) the budget for the evaluation is limited; and/or (4) the evaluation is primarily for program improvement rather than accountability (see Part 4).

> **Primary Evaluator**
> The lead of an evaluation project

Alternatively, external evaluators are hired from outside the organization/institution to consult as an unbiased evaluation expert when: (1) an outside perspective is essential; (2) the program is large or complex; (3) there is a flexible budget for the evaluation; and/or (4) the evaluation is primarily for accountability rather than improvement (see Part 4). By having determined the research questions, and scope and capacity for the evaluation, primary evaluators can develop a work plan so that they can prioritize what needs to be done, and then find or build capacity for satisfying other possibilities on the program evaluation wish list.

Once the primary evaluator has been selected, the next consideration is whether and to what extent other stakeholders will be involved in the evaluation. Stakeholders' participation can vary widely across a spectrum, largely dependent on the research questions, the scope of the project and how the results will be used. Small programs with limited resources might have a relatively short list of stakeholders to consider, whereas larger or more complex programs may have a lengthy list of stakeholders worth involving. If the evaluation is too difficult to be coordinated by a single person, an evaluation team consisting of applicable stakeholders might be created to support the primary evaluator. The types of stakeholders involved will also depend on the type of organization/institution in which the evaluation is taking place (e.g., community groups, colleges or universities). Figure 5 lists some common stakeholders that may be considered for inclusion within evaluations of access- or retention-related programs.

> **Stakeholders**
> Individuals or groups who have a vested interest in the process and/or results of an evaluation project
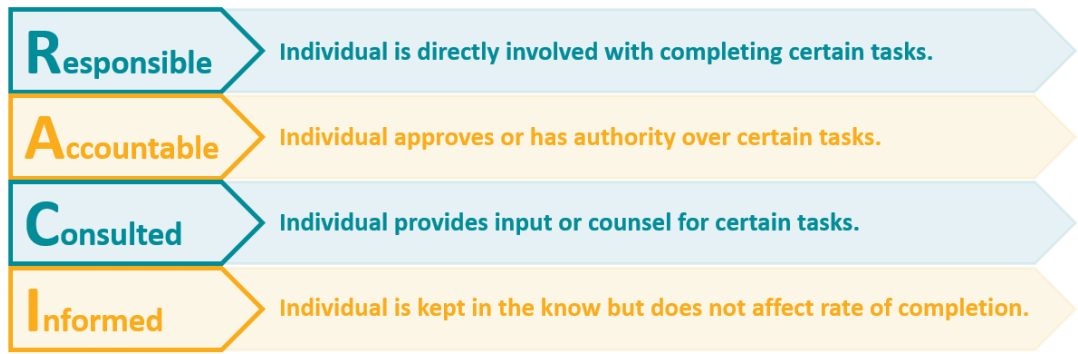
Figure 5: Potential Stakeholders to Consider for Inclusion in Evaluating Large or Complex Programs

Regardless of the size of a program or the corresponding evaluation team, it will be helpful to identify the role(s) for each stakeholder involved by categorizing their expected contribution towards important tasks.[7] A common approach for characterizing stakeholders' roles is shown in Figure 6.

---

7 For any stakeholder that will not be responsible, accountable, consulted or directly informed about an evaluation (e.g., taxpayers, future participants), these perspectives may still be considered when evaluators assess whether the plan for evaluation implementation and reporting is ethical for all parties (see Part 10).

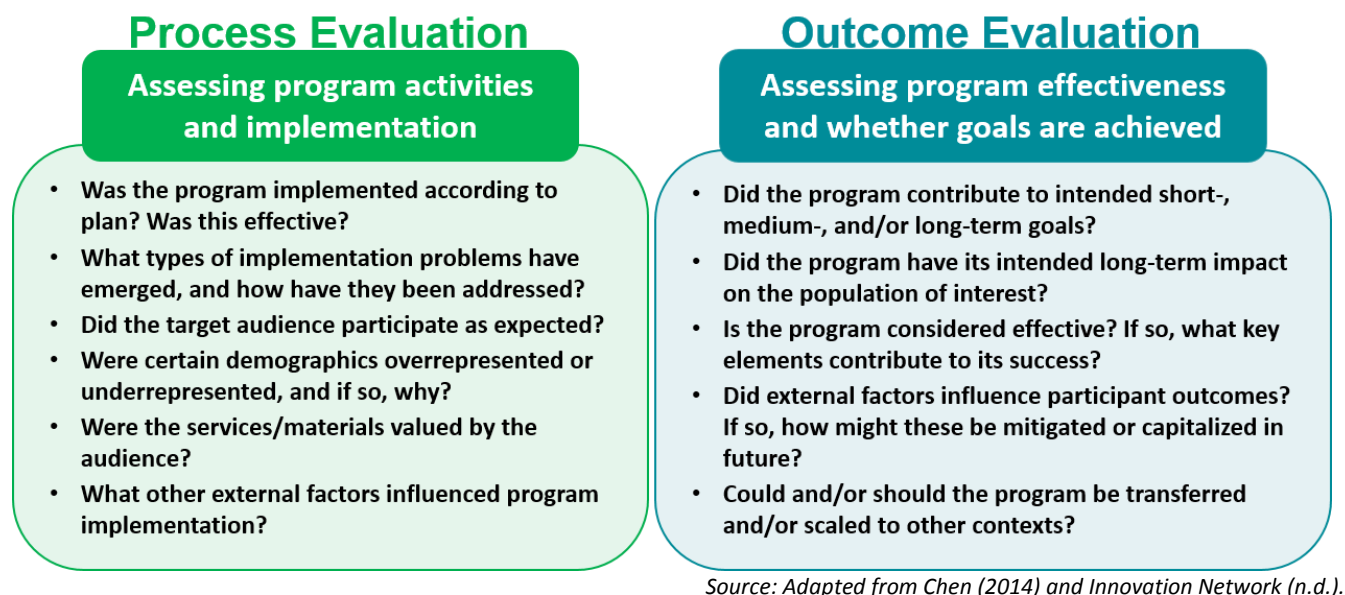**Figure 6: Characterizing Stakeholder Involvement Using the RACI Acronym.**

Responsible	Individual is directly involved with completing certain tasks.
Accountable	Individual approves or has authority over certain tasks.
Consulted	Individual provides input or counsel for certain tasks.
Informed	Individual is kept in the know but does not affect rate of completion.

The process of defining roles for evaluators and stakeholders is more an art than a science. Using the research question(s) and evaluation scope as justification, evaluators should rationalize stakeholders' inclusion or exclusion in the evaluation project. When stakeholders are engaged, this can increase the likelihood of capturing their expertise, facilitating their acceptance of results and inspiring their action for program improvement. However, maximizing stakeholder engagement is not always feasible or beneficial depending on the nature of the project. Sometimes, it can be detrimental to have too many cooks in the kitchen, and evaluators should avoid involving stakeholders unnecessarily or too early in a project. Ultimately, whether stakeholders are purposefully included or excluded, evaluators should strive for communication that is clear, consistent and respectful.

Part 4: Selecting an Evaluation Approach

After establishing research questions, and choosing the evaluator or evaluation team, the next step might be to establish an overarching evaluation approach. The approach taken to a program evaluation is concerned with "what" is being evaluated, and should be revealed by the research question(s). Most often, the evaluation approach is categorized as being either a process evaluation or an outcome evaluation (Figure 7).[8] For evaluations with multiple research questions, evaluators may take a hybrid approach, incorporating components of both process and outcome evaluation. Distinguishing which questions lend themselves to each type of approach is very important, as this will dictate other choices to be made moving forward.

Figure 7: Comparing Process and Outcome Evaluations and Associated Research Questions

Process Evaluation

Assessing program activities and implementation

- Was the program implemented according to plan? Was this effective?
- What types of implementation problems have emerged, and how have they been addressed?
- Did the target audience participate as expected?
- Were certain demographics overrepresented or underrepresented, and if so, why?
- Were the services/materials valued by the audience?
- What other external factors influenced program implementation?

Outcome Evaluation

Assessing program effectiveness and whether goals are achieved

- Did the program contribute to intended short-, medium-, and/or long-term goals?
- Did the program have its intended long-term impact on the population of interest?
- Is the program considered effective? If so, what key elements contribute to its success?
- Did external factors influence participant outcomes? If so, how might these be mitigated or capitalized in future?
- Could and/or should the program be transferred and/or scaled to other contexts?

Source: Adapted from Chen (2014) and Innovation Network (n.d.).

Evaluators may also want to clarify the purpose of the program evaluation (i.e., how the results will be used). Again, the purpose of the evaluation should be clearly extrapolated from well-written research questions, which may be sorted into one of two meta-categories: constructive or conclusive.

Constructive Evaluations
Evaluations undertaken to improve a program
(sometimes known as formative or learning-focused)

Conclusive Evaluations
Evaluations undertaken to prove a program's effectiveness
(sometimes known as summative or accountability-focused)

---
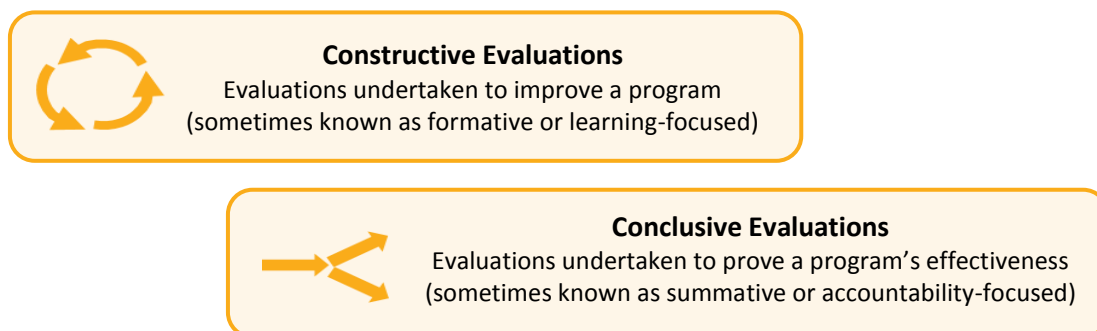
8 Some evaluators may prefer the term "implementation evaluation" to process evaluation.

Although it may seem as though process evaluations would be associated with constructive questions, and outcome evaluations with conclusive questions, this is not always the case. Findings from either a process or outcome evaluation can be used for either constructive or conclusive purposes. And evaluations with several research questions might fall into another hybrid category: having some questions that are constructive and some that are conclusive. What's important here is that evaluators recognize how to distinguish the goals of each research question, as this will be crucial for selecting a study design (Part 5).

> **Case Consideration: Examples of constructive outcome and conclusive process evaluations**
> Let's compare a few different approaches for evaluating the application subsidy program in Figure 3. A constructive outcome evaluation might survey subsidy recipients one year later to determine whether they are attending PSE (a medium-term outcome) and whether they have any constructive feedback about the marketing of the program for future student cohorts. Alternatively, instead of waiting for a year to pass, a conclusive process evaluation might be undertaken to evaluate the effectiveness of the marketing strategy during the initial implementation of the subsidy program. If such an evaluation were to find that the marketing scheme was completely ineffective, and that most students from the target population did not apply for the subsidy, stakeholders may cut the program. This would allow resources to be re-allocated to better serve the target population in a timely manner.

Besides being constructive or conclusive, evaluations could also have a primary focus that does not fit neatly into these two meta-categories. Such evaluations might be focused on building trust and legitimacy with stakeholders, contributing to a broader evidence base, or assessing the potential to scale up or transfer a program to other contexts, etc.

> **Case Consideration: Hamilton Community Foundation's ARC project using a multiphase hybrid evaluation**
> To evaluate a pilot program at the middle school level for students facing multiple PSE access barriers, Hamilton Community Foundation is using a multiphase hybrid evaluation approach. Since the program is new and has an emerging theory of change, the evaluation begins with a constructive process approach to understand the core program components and the goals and objectives for the pilot. Once these questions have been answered, Hamilton Community Foundation intends to take a more conclusive approach to outcomes-focused evaluation to determine the program outcomes (both intended and unintended) on participants, as well as the scalability of the program for other contexts. The use of multiple evaluation approaches and questions necessitates a mixed-methods study design.

Ultimately, no matter how complex an evaluation, it is important to align each research question with (1) an investigation of either the process(es) or the outcome(s), and (2) a distinct purpose, be it constructive, conclusive, or some other primary driver. Without dwelling on the semantics of hybrid types of evaluations, or the many unique approaches to evaluation that exist, we hope that defining what portion of the program is being evaluated and how results will be used will help evaluators establish their overarching evaluation approach. By establishing these parameters, selecting an appropriate study design should be much more straightforward.

Part 5: Choosing a Study Design

After finalizing research questions and establishing an evaluation approach, the next step is to select a suitable study design. Certain research questions will require (or preclude) the use of particular study designs. The study design is very closely tied with the evaluation's purpose, so this choice will depend on whether the results are to be used for constructive (improvement-focused) or conclusive (accountability-focused) purposes. Furthermore, outcomes evaluations will generally require the use of experiments or quasi-experiments, whereas process evaluations can use any design, but most commonly align with non-experimental designs.[9] The three designs are described here:

Experimental Designs assess a program's unbiased effects by having participants randomly assigned to either a control or intervention group to generate equivalent groups for comparison. Each group is commonly assessed at least twice: before and after the program or intervention.

Intervention Group	**Control Group**
Individuals who participate in the program (random assignment depends on study design)	Individuals randomly assigned to not participate in the program

Also known as randomized control trials (RCTs), experiments have historically been considered the gold standard in program evaluation because they are thought to provide the most credible evidence of a causal relationship between a program and an outcome.

> **Case Consideration: Mohawk College's ARC project using a randomized control trial**
> This study used an experimental design to measure the effects of proactive advising, a student service whereby staff support students in developing personal and academic goals throughout their first year. In this study, approximately 4,500 students were assigned to a control group, or one of two intervention groups (to either receive group advising or one-on-one advising). This design allowed researchers to perform statistical analysis to compare the effects of group advising and one-on-one advising to baseline data generated from the control group (i.e., no advising).

Quasi-experimental Designs assess a program's effects by analyzing participants from comparison and intervention groups. Both intervention and comparison groups are commonly assessed twice: before and after the program. Evaluators using quasi-experimental designs should be mindful that advanced statistical analysis is usually required for this method to control for differences between the groups, which in turn isolates program effects.[10]

Comparison Group
An adapted version of a control group that avoids randomization by selecting a group of non-participants as similar as possible to the intervention group to minimize biases

---

9 For more information on selecting a study design, see the Pell Institute's website here.
10 Even though the comparison group has not received the intervention, they are still assessed at the same time-points as the intervention group in quasi-experiments.

> **Case Consideration: University of Toronto Scarborough's ARC project using a quasi-experiment**
> This study will use a quasi-experimental design to measure the impact of the first-year orientation program, *Get Started* (described in Figure 4). Using propensity score matching, a statistical technique that creates a comparison group by matching demographic characteristics to the intervention group, researchers will assess whether program participation impacts participants' retention and academic success, especially focusing on the effects for first-generation and low-income students.

Non-experimental Designs assess how a program functions by primarily analyzing the intervention group. Some designs may track participants over time while others may only assess participants once, after the program or intervention. Rather than assessing program effects, non-experiments are often used as a descriptive tool to understand elements of a program. For some contexts, this study design is the most realistic, especially when creating a control/comparison group is impossible, unethical or unaffordable. Even so, efforts to find some means of comparison for program participants' outcomes (e.g., benchmarking against standards or targets, or using census data) can be used to provide a useful point of reference.

> **Case Consideration: Success Beyond Limits' ARC project using a non-experiment**
> To evaluate their youth-led mentoring program, Success Beyond Limits (SBL) is performing Youth Participatory Action Research. This approach will allow SBL alumni to conduct interviews with fellow alumni to better understand the role of the SBL program on PSE trajectories and outcomes. This approach can be beneficial when looking to gather data retroactively, as it can gather data from participants even after they have participated in the program.

Besides having a general understanding of what each of these three study designs entail, evaluators will also have to keep the following considerations in mind:

Feasibility Concerns	Validity and Reliability Expectations	Stakeholder Preferences
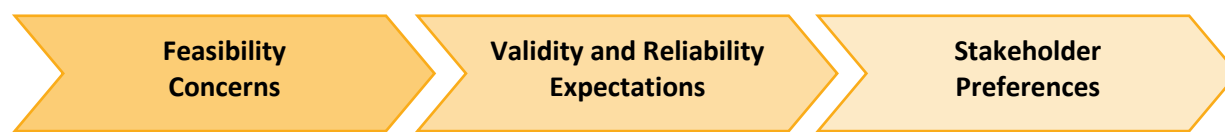
In terms of feasibility, RCTs and quasi-experiments can be quite resource intensive to implement. Besides the time and budget required, these designs usually necessitate experienced evaluators with ample expertise in data analysis. Additionally, ethical considerations can be a significant obstacle, as the randomization of participants to certain programming may not be justifiable (see Part 10). An exception to this issue is programs for which there is a limited evidence base, as RCTs tend to be more easily justified to learn about a program's effects for the first time.

In terms of validity and reliability, RCTs have historically tended to shine, with quasi-experiments a close second. However, the inherent validity of RCTs has been called into question when dealing with program evaluations. Although it is well established that traditional experiments have the highest internal validity of any study design, this design has comparably lower external validity. This is because external factors normally embraced and/or mitigated by program staff to ensure a program's success within its real-world context are eliminated in the experimental study design. Consequently, the results from RCTs are not always generalizable to the very context in which the program is normally situated.

> **Internal Validity**
> The extent to which a study is measuring what it intends to

> **External Validity**
> The generalizability of results for other populations, settings, variables and measurements

The final consideration when selecting a study design might involve a balancing of stakeholder preferences, if applicable. Regardless of other considerations, evaluators may be limited by stakeholders' preconceived views about the rigour of certain study designs.

What's most important is that evaluators use the research question and associated evaluation approach to guide their choice of an appropriate study design. By acknowledging how the advantages and disadvantages of the selected study design align with the evaluation goals, they can be confident in justifying their choice moving forward.

B Performing an Evaluation

Part 6: Collecting the Data

Once a study design has been selected, it's time to plan how the data will be collected. Referring back to Table 1 (Part 2), evaluators might categorize the variables required for each research question as either quantitative or qualitative. If only one type of variable is required, evaluators will be using solely quantitative or solely qualitative methods to answer said question. If, however, evaluators need both types of variables to answer research questions, a mixed-methods approach will be used.

Mixed-methods approaches are gaining significant traction in the program evaluation sphere due to their ability to overcome the disadvantages associated with either singular source of data. Mixed-methods approaches that use a variety of data sources (i.e., variables) and collection techniques (e.g., surveys, interviews) are more valid and reliable. Not only do the multiple data sources contribute to data triangulation, but they can inform results in different ways, and thus increase the depth of analysis and confidence in reporting.

> **Data Triangulation**
> The validation of data through cross verification from two or more sources/techniques

While we could sort specific data collection techniques according to the type of data they produce, this is made difficult by the fact that many techniques can be used to produce multiple forms of data (i.e., surveys can be either open- or close-ended, providing either qualitative or quantitative data, or both). For this reason, data collection techniques are also often sorted according to the source of data, as follows:[11]

- **Reviewing existing documentation** — examine databases, official statistics, archives, financial records, literature review, project records, previous evaluation results, etc.
- **Observation** — participant or non-participant observation, site visits, review photos or videos, etc.
- **Gathering information from individuals** — interviews (structured or unstructured),[12] diaries/logs, surveys (open- or close-ended), etc.
- **Gathering information from groups** — focus groups, concept mapping, debates, etc.

Within each category listed above, there will be several considerations that will lead to the selection of the most appropriate data collection technique. Analyzing an institution/organizations' internal documents can provide evaluators an understanding of their resources, values, processes, priorities and concerns to identify strengths and inefficiencies. Analysis of public records can be essential for making large-scale comparisons among groups and communities. Observing programs in action can provide evaluators a holistic perspective about the context in which the program is situated and can be useful for exploring issues that staff or participants are unaware of, unwilling or unable to discuss in an interview. When gathering information from individuals or groups, there are several advantages and disadvantages to consider. These are summarized in Table 2.

---

11 For more information on data collection techniques, see BetterEvaluation's website here.
12 For both interviews and focus groups, evaluators should consider how they will reliably capture detailed information. If it is not feasible to take detailed notes, it is recommended that evaluators use a recording device, especially if there is an interest in directly quoting participants.

Table 2: Comparison of Techniques for Gathering Information from People

Description	Pros	Cons
Close-ended Surveys Collects info from large groups about topics that are neither complex nor sensitive using clearly defined questions and predetermined answers	• High reliability • Efficient data collection • Simple data analysis • Low cost	• Difficult to construct • Self-report can cause bias • Impersonal and lacking context • Limited validity • Not suitable for complex or sensitive topics
Open-ended Surveys[13] Collects info from moderately sized groups prompting open responses to clearly defined questions on topics that can be somewhat complex or sensitive	• Moderate validity • Efficient data collection • Moderate cost	• Difficult to construct • Self-report can cause bias • Impersonal • Difficult data analysis • Limited reliability
Interviews Collects rich contextual or behavioural info about feelings and/or opinions from a small sample of knowledgeable people	• Produces rich data • Allows interpersonal contact • High validity • Suitable for complex/sensitive and high-status respondents	• Requires trained interviewers • Lengthy data collection • Difficult data analysis • Limited reliability • High cost
Focus Groups Collects relatively non-sensitive info from groups of 8–12 people when group interactions can be harnessed to stimulate rich responses	• Produces rich data • Allows interpersonal contact • High validity • Moderately efficient data collection • Moderate cost	• Requires trained interviewers • Difficult data analysis • Limited reliability
Observation Collects data on a wide range of behaviours to learn about issues that program staff or participants may be unaware of, or unable or unwilling to discuss	• Produces direct and holistic data about behaviours • Provides good sense of context to evaluator • Occurs in natural, unstructured settings	• Requires trained observers • Potential for biased data (e.g., wary participants, selective perception) • Difficult data analysis • Limited reliability • High cost

Source: Adapted from National Science Foundation, 2002; Baker, n.d.

Upon reaching a decision about how all the data for an evaluation is to be collected, evaluators can create a data work plan to keep the data collection and analysis process organized. Work breakdown structures are a commonly used tool in project management that can be used to identify deliverables and associated leads for each task to be completed during the evaluation.[14] Designing a schedule will also be essential to ensure the team has sufficient time for data analysis and report writing. A Gantt chart is a commonly used tool for visualizing timelines for complex projects in which multiple individuals are tasked with various activities and corresponding deadlines.[15] These types of tools can support other aspects of evaluation planning concerning the budget, and resource allocation (see Part 11).

---

13 Both close- and open-ended questions can be effectively used in one survey tool.
14 For more information on work breakdown structures, see Project Management Docs' website here.
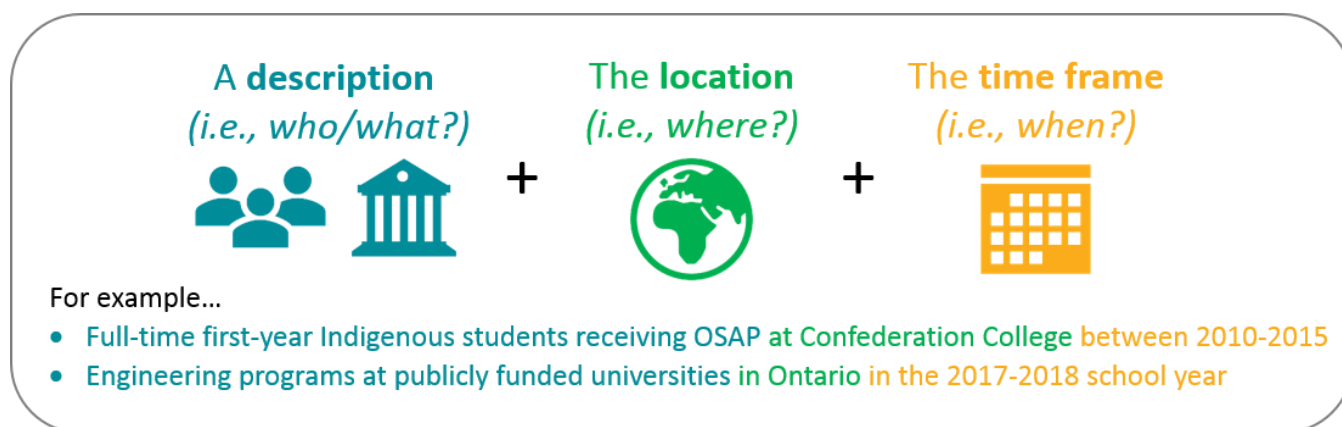15 For more information on Gantt charts, see Project Manager's website here.

Part 7: Sampling the Population

Once the study design and data collection methods have been established, the next step is to further define who is being evaluated and why. Evaluators should use research questions and the scope of the evaluation as the basis for establishing their specific population of interest. To define a population of interest, evaluators should include the three components described in Figure 8.
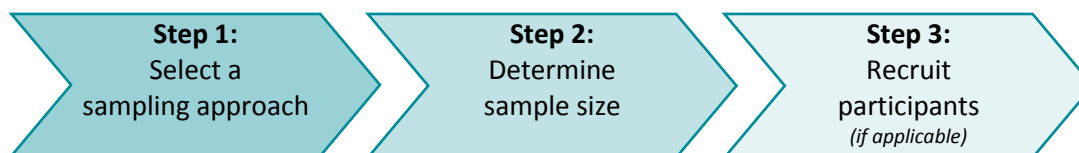
> **Population of Interest**
> The group to be learned about; made up of individuals or any other unit of data (e.g., institutions, organizations, records, etc.)

Figure 8: Defining a Population of Interest

A **description** (i.e., who/what?)

+

The **location** (i.e., where?)

+

The **time frame** (i.e., when?)

For example…
- Full-time first-year Indigenous students receiving OSAP at Confederation College between 2010-2015
- Engineering programs at publicly funded universities in Ontario in the 2017-2018 school year

Once the population of interest has been defined, evaluators will determine which cases (i.e., members or units) could feasibly be recruited to participate in their study. If one's population of interest is quite small, or if the research question requires that all program participants be included, evaluators can include all cases in their study. For example, if an evaluator wants to examine the impact of a particular training session for residence dons in female-only residence buildings, they could include all dons who participated in that training program in their study. For these types of evaluations that include one's entire population of interest, evaluators can skip to Step 3 below. Alternatively, if evaluators have large populations of interest, or if they want to study a specific subset of their population, they can follow these steps:

Step 1: Select a sampling approach → **Step 2:** Determine sample size → **Step 3:** Recruit participants *(if applicable)*

Step 1: Select a Sampling Approach

To select a subset of one's population, evaluators can choose between probability and non-probability sampling.[16] The choice of sampling technique will depend on the goal(s) of the evaluation, the research design, and the feasibility regarding capacity, time, budget and ethical constraints.

---

16 For more information on sampling methods, see BetterEvaluation's website here or the Pell Institute's website here.

> **Probability Sampling**
> Selects a subset using random or quasi-random techniques to make generalized inferences about the larger population of interest

Probability Sampling. To use probability sampling, evaluators must have a way to identify each case in the population that is accurate, exhaustive (i.e., including the entire population of interest), up to date, free of duplicates, and absent of any pattern in how member/unit names appear in a list. Then samples can be selected using:

- **Random sampling** — Each member/unit of the sample is chosen randomly from the population and has the same likelihood of being chosen for the sample.
    - e.g., A computer program selects 30 students from a list of all psychology majors attending Ryerson University.
- **Stratification sampling** — First the population is split into distinct groups to maximize representation, then cases are randomly selected from each group.
    - e.g., The names of all psychology majors enrolled in universities across Ontario are first sorted by university, and then a computer program selects 30 students from each institution.

Although probability sampling is ideal for making inferences about one's population of interest, it is not always necessary (e.g., if learning about individual cases is the priority) or feasible (e.g., it may be expensive, or difficult to get a full list of all the members in a population).

> **Non-probability Sampling**
> Selects a subset using non-random techniques, which may or may not allow inferences to be made about the larger population of interest

Non-probability sampling. If the aim of an evaluation is to learn about cases for some purpose other than generalizing, then non-probability sampling is appropriate. Non-probability samples can be selected using:

- **Purposive sampling** — selecting a subset based on one or more pre-determined characteristic(s) (e.g., selecting cases that are typical, critical (i.e., important in some way), outliers or aligned with a certain criterion or pattern
    - e.g., Recruiting third-year psychology majors attending Ryerson University who have intentions of pursuing graduate studies
- **Convenience sampling** — selecting samples which are readily available (e.g., selecting volunteers or using "snowball" sampling)
    - e.g., Having a teaching assistant recruit Ryerson students from their tutorial section for a 300-level psychology course

Purposive sampling allows evaluators to select cases that have inherent meaning to the study so that they can make specific or analytic inferences about the larger population of interest. Convenience sampling does not allow for credible inferences to be made about the larger population. In both types of non-probability sampling, biases in the sample will prevent generalizations to the entire population of interest, and are thus important to acknowledge when describing the findings.

Step 2: Determining Sample Size

When deciding on sample size, there will be several trade-offs between breadth and depth, that take into account the type of data being collected (quantitative versus qualitative), the evaluator's intentions for generalizing the findings and the feasibility. Since qualitative data can be more resource intensive to collect, evaluators may be limited to smaller sample sizes, necessitating more depth than breadth to be feasible. On the other hand, for evaluators hoping to be able to say something about their entire population, depth will often be sacrificed for breadth. In these situations, evaluators must capture a large enough sample to give an accurate snapshot of the entire population so that generalizations can be made. In some instances, an evaluator can use a power analysis to help determine what sample size is suitable to accurately detect the effect they are interested in.[17] When evaluators are not concerned with making generalizations, they may rationalize their choice in sample size by (a) selecting the sample size in advance, (b) sampling to the point of redundancy/saturation, or (c) using emergent sampling to add cases as applicable over time. Evaluators may even choose to study a single case if that suits their study design.

Step 3: Recruiting Participants

To collect data prospectively from individuals, recruitment techniques are an essential consideration, especially when individuals' participation in a study is completely optional. This step can often take much longer than anticipated, but is certainly worth the time and effort. Stakeholders may be worth consulting again at this stage as they could potentially act as liaisons with participants and/or provide guidance about recruitment techniques. Recruiting participants can range in difficulty depending on the size of one's population of interest, the relationships between members and the sampling approach. Advertising is a common way of spreading the word about the participation opportunity, whether this is done in person, using print or online. Incentives are also a useful strategy to increase participation when feasible, but these have associated ethical considerations (see Part 10). Evaluators should also anticipate and/or react to non-response rates, especially when a strict minimum sample size is required. To mitigate this, evaluators might increase the number of participants they initially recruit accordingly. Lastly, evaluators should also be sure to acknowledge how their recruitment strategies could bias the results when reporting the findings.

Once evaluators have defined their population of interest, determined their sampling approach, selected their sample size, and recruited their participants, they can finally initiate their study design and begin collecting the data to inform their evaluation results.

---

17 For more information on selecting sample sizes, see Taylor-Powell (1998)'s resource here.

Part 8: Managing, Storing and Analyzing Data

There are several important considerations when it comes to storing, managing and analyzing program evaluation data. It is recommended that evaluators include these details when developing their data work plan, prior to initiating their evaluation so that tasks can be completed in a timely manner. The processes for inputting, organizing, storing and cleaning the data prior to analysis should be clearly laid out for the team member tasked with each activity. Depending on the capacity of the evaluation team, it may be most efficient to hire an external consultant when expertise is lacking in-house.

Data Management: The processes for inputting, organizing and cleaning the data prior to analysis should be clearly laid out for each member of an evaluation team. This is especially important when multiple team members will be handling the data, as a strict protocol should be followed to ensure that all data is always handled in the same way. For instance, if the same survey is given both in person and online, the evaluation team will want to have a plan for manually coding and inputting data from the physical copies to match the outputs from the online version. As previously mentioned, work breakdown structures can be very helpful in allocating tasks and deliverables for team members during this phase of work.

Data Storage: The major consideration regarding data storage pertains to privacy. The level of security required for data storage will depend on the sensitivity and confidentiality of the information, which should be outlined within ethics protocols (see Part 10). Establishing a secure and reliable strategy for storing information *before* it has been collected is essential so that data is never left vulnerable to corruption, destruction or misuse.
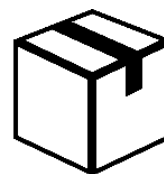
Data Analysis: Unlike many other components of evaluation, the skills required for data analysis are very specific and sometimes hard to come by. When this expertise is not available through one's evaluation team or larger network, it may be more cost effective to hire an external data analyst than to train someone in-house for a single project. The data analyst may have specific expertise in either quantitative or qualitative analysis, as these methods will differ substantially.[18] The value of these skills should not be underestimated, as data analysis will very directly affect the credibility of one's results. In either case, data cleaning is an important first step whereby potential errors and/or inconsistencies in the data are removed or corrected to ensure accuracy of the data set.

Quantitative Data Analysis: This is the process of turning raw numbers into meaningful data through the calculation of statistics.[19] Descriptive statistics (e.g., the mean, median, maximum and minimum, etc.) allow analysts to summarize and examine raw data, whereas inferential statistics allow analysts to make inferences about the population of interest. Analysts might take the following steps to analyzing their quantitative data:

Step 1: Inputting Data onto an Analytic Software
Analytical software is used to collect, store, organize, visualize and interpret one's data. The particular software to be used should be selected before data collection based on the type of statistics to be calculated. For example, programs like Microsoft Excel will be sufficient to calculate descriptive statistics, whereas programs like SPSS or STATA may be better suited for calculating inferential statistics.

---

18 For more in-depth information on methods of data analysis, see Cozby & Rawn (2016).
19 For more information on quantitative analysis, see the Pell Institute's website here.

Step 2: Selecting Appropriate Descriptive Statistics

The two most important descriptive statistics to summarize data sets are central tendency (e.g., the mean, median, and/or mode) and variability (e.g., the range and/or standard deviation). Choosing the most appropriate statistics to summarize one's data depends on each variables' scale of measurement (i.e., the ways in which variables/numbers are defined and categorized). There are four scales of measurement, which are summarized in Table 3 below. Correlation coefficients can also be used to describe the nature (e.g., positive, negative or neutral) and strength (e.g., strong or weak) of the relationship between the two variables of interest.

Table 3: Scales of Measurement

Scale	Defining Features	Example	Appropriate Statistics
Nominal	Categories or numbers that are used as identifiers or names	• Classifying sex or race • Intervention /control group	• Mode • Percentages
Ordinal	Categorical data that is ranked in order but with no particular value for the differences between rankings	• Social class • Birth order	• Mode • Median • Percentages
Interval	Numeric properties are literal and the difference between numbers of the scale are equal in size	• Year • Likert scales	• Mean • Standard deviation
Ratio	Numeric properties are literal with a true zero and the difference between numbers of the scale are equal in size	• Time to complete a test • Age • Frequencies of behaviours	• Mean • Standard deviation

Step 3: Selecting Appropriate Inferential Statistics

Unlike descriptive statistics, there are many different inferential statistics that can be calculated to learn about one's population of interest. While an exhaustive list of all possible analyses is beyond the scope of this manual, a few common techniques include:

- **Regression** — an extension of correlation used to examine and predict the relationship between two or more variables of interest
- **T-test** — used to determine whether a statistically significant difference exists between the means of two groups (e.g., an intervention and control group)
- **Analysis of Variance (ANOVA)** — similar to the t-test, used to determine whether a statistically significant difference exists between the means of three or more groups
- **ANCOVA** — similar to ANOVA but also allows the evaluator to control for confounding variables when determining statistically significant differences between the means of groups

> **Case Consideration: Using quantitative analysis for Mohawk College's ARC project**
> This study assessed the impact of proactive advising using regression analysis to determine whether and to what extent their intervention could predict retention. Results of the regression demonstrated that proactive group advising sessions were more effective than proactive one-on-one advising sessions, and that only males saw a statistically significant improvement in retention rates from group advising.

Qualitative Data Analysis is the process of finding meaning from narrative and/or anecdotal data, often collected through interviews, focus groups, open-response surveys, observations or document analysis.[20] This process is less standardized than quantitative data analysis, but can provide very rich information about the nature of a program and the experiences of participants. On a practical note, the data can be coded manually, using programs specifically designed for qualitative analysis (e.g., NVivo), or via some combination of the two.

If the same member of an evaluation team is tasked with both data collection and analysis, these processes may not occur in isolation. With the exception of structured interviews (i.e., consultations with pre-determined questions), qualitative data collection and analysis can be an iterative process whereby analysts make sense of the information while it is being collected. Although the iterative nature of this process may introduce bias, it can be beneficial as it allows analysts to identify emergent themes or questions to probe further in remaining consultation sessions. Once all data has been collected, analysts should use research questions as a framework to perform data reduction (i.e., simplifying and condensing the original data into codes to focus on what is most meaningful to the project). The same data can be analyzed from multiple perspectives depending on the research questions being addressed.[21]

Once the analysts have reduced their data to a manageable level, the coded data can be analyzed for meaning in a variety of ways. Such meaning can either evolve directly from the research questions or emerge naturally as the study is conducted. Data is most often coded for both content (e.g., identifying key words/phrases) and for themes (e.g., positive versus negative sentiments). It is also important to note the relative frequency with which different topics are discussed and the intensity with which they are expressed. As mentioned previously, data triangulation can be used to increase confidence in reporting, as the use of multiple methods/sources can strengthen evaluation results. Data triangulation still applies even if both methods of data collection are qualitative, such as when observational data can be corroborated by interview data. If, however, evaluators find discrepancies between the various methods of data collection (e.g., data from observations are not in agreement with data from interviews), then they must use their expertise to weigh the evidence and/or investigate further to understand the differences. Stakeholders may be involved again at this stage to provide any necessary details or nuances for the analysis and to discuss any surprises or unintended results found in the data prior to reporting.

> **Case Consideration: Using qualitative analysis for Queen's University's ARC project**
> Queen's sought to compare the in-person and online versions of their first-year transition program, "Q-success," and the differential impacts among historically underrepresented students. Analysis of qualitative responses from the Thriving Quotient (TQ) questionnaire found a decrease in thriving for students who attended the program, and for students who did not, suggesting that the TQ was not an effective tool to document the impact of this program. Students' written responses revealed that scheduling conflicts and other commitments deterred them from participating in the full program.

---

20 For more information on qualitative analysis, see the Pell Institute's website here.
21 If an external data analyst who is unfamiliar with the program is tasked with data reduction, someone involved with the program might assist the analyst at this stage.

Part 9: Reporting Results

Once all the data has been collected and analyzed, it is time to think about how the evaluation results will be reported to one's audience. When it comes to reporting, the one principle that matters above the rest is relevance. Evaluators should think about how findings can be framed as relevant to the range of stakeholders being served by the evaluation. This will likely require a balancing act, as evaluations are often designed quite holistically with a number of interrelated and integrated elements to report on. Most stakeholders will have a vested interest in a particular aspect of an evaluation; and they will want these aspects highlighted appropriately when results are being reported. With this in mind, it is recommended that evaluators think about how results will be reported to stakeholders during the planning stages of evaluation. Depending on the type of partnership agreement(s) with various stakeholders, it may be necessary to ensure that all applicable parties are informed about how results will be disseminated, and who is responsible for doing so. The ways in which results are reported may be dependent on one's ethical protocol (see Part 10).

It is also important to discuss in advance how the findings might affect the actions of different stakeholders moving forward. In particular, agreements should be reached about how null or negative results will be reported. An important concern often expressed by stakeholders is the potential that evaluation findings could be misrepresented, taken out of context or misunderstood. As such, it is paramount that evaluators consider how, when, with whom and for what purpose results are reported. Some stakeholders may want a nuanced report that tells a story, whereas others may prefer concise reporting that focuses mostly on measuring outcomes numerically. Managing perceptions is also important for programs to maintain funding, especially when public funding is at stake, as taxpayers might have strong opinions about these investments.

> **Case Consideration: Considering stakeholders' perspectives for reporting evaluation results**
> If we evaluated the application subsidy program from Figure 3 to understand short-, medium- and/or long-term program outcomes, there might be several key stakeholders with unique vested interests. Low-income students and their advocates (e.g., parents, student groups and associations) likely care about reducing financial barriers to PSE. While evidence from the evaluation could suggest fee subsidies do not actually support an increase in attendance among low-income students, keeping the option to apply without paying fees might be a point of contention. Program staff might want to understand the impact of application subsidies isolated from other interventions they may be delivering to students (e.g., one-on-one support, outreach and engagement). And funders/donors might be most concerned with the return on investment, looking to medium- and long-term outcomes of how many low-income students accessing the subsidy attend and graduate from PSE.

Given the different interests and investments at play, it is important for evaluators to consider whether different approaches will be taken for reporting to various audiences. Results may be reported in a range of formats (e.g., written reports, presentations, videos, posters, memos, etc.) and should be adapted to each target audience. This will be a matter of emphasis, using judgment to identify what is important to share, with whom, at what time and for what purpose. Similarly, the depth and frequency of reports will need to be identified early in the project. Contracts with funders will likely detail reporting requirements, but it will also be important for evaluators to weigh the needs of all stakeholders when reporting results. It is recommended that evaluators prioritize keeping a good relationship with funders/donors, while also keeping them apprised of the needs of all other stakeholders, to ensure the integrity of reporting and the evaluation as a whole.

C **Additional Factors to Consider for an Evaluation**

Part 10: Meeting Ethical Standards

A critical consideration for evaluations is that they meet the necessary ethical standards. This is especially important for research/evaluations involving human participants. Evaluators should ensure that all activities throughout the evaluation, from selecting and recruiting participants, to collecting and storing the data, to interpreting and reporting findings, are completed with ethical standards in mind. Depending on the organization or institution, there may be specific internal processes or guidelines that evaluators must adhere to. In the context of PSE, all postsecondary institutions have a Research Ethics Board (REB) approval process that may need to be followed.[22]

> **Research Ethics Board**
> A panel of experienced researchers and/or administrators who thoroughly review the purpose of, and methodological approach to, any research/evaluation project involving human participants

REBs exist to ensure that research meets high scientific and ethical standards that respect and protect participants in adherence with the Canadian Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans.[23] The Tri-Council statement requires that "all research involving humans be conducted in a manner that is sensitive to the inherent worth of all human beings and the respect and consideration that they are due." Any potential harms anticipated for the study participants (be they physical or psychological) can be justified in terms of the potential benefits to them, to others and to society as a whole.[24] Researchers who intend to engage in research with human participants are advised to take the TCPS2 Tutorial (and may be required to submit their certificate of completion with an application).

Typical components of an REB application often include:

- ✓ Background/study description
- ✓ Research question
- ✓ Study design/proposed methodology
- ✓ Participant descriptions
- ✓ Recruitment plan
- ✓ Consent process

- ✓ Data security
- ✓ Confidentiality and protection of anonymity
- ✓ Dissemination
- ✓ Conflict of interest
- ✓ Potential risks
- ✓ Potential benefits

Besides postsecondary institutions, evaluators from other types of organizations may still need to get REB approval or follow another mandated ethics approval process. For any organization that doesn't have a research ethics approval process in place (e.g., community groups), evaluators might review a typical REB process to outline ethical guidelines that the project can adhere to.

Among the ethical considerations covered within a typical ethics approval process, there are a few points especially relevant for access- and retention-related programs that warrant further discussion.

---

22 School boards also have research ethics protocols to be referenced when research is conducted with students in grades K–12.
23 For those interested in a more applied approach to the guidance in the Tri-council Policy Statement, consider the Government of Canada's Course on Research Ethics.
24 For more information on the potential harms and benefits of a study, see the National Research Council of Canada's (2012) website here.

The first consideration is about receiving informed consent from participants. All participants in an evaluation should be informed about the purpose of the evaluation, what their participation entails, all possible risks of participation and how their information will be protected. It should also be clear to participants that they may withdraw their consent to participate at any point throughout the study. If potential participants' autonomy to give consent is either diminished or impaired, participants may require extra protection or consent from a third party who is entrusted to make decisions for the participant. Requiring informed consent allows evaluators to communicate to participants what they need to know about the evaluation in a way that is clear and concise, using suitable language so that it may serve as a basis for any questions or discussion about their consent.[25]

The next ethical consideration is regarding the use of incentives to increase sample size. According to the Tri-Council statement, incentives are neither recommended nor discouraged, and the onus is on the evaluator to justify the use of incentives. Whether incentives are financial or some other form of reward, evaluators must be cautious that the incentives are appropriate in type and amount. Incentives should not be so great that they could be a form of coercion for reckless disregard of the associated risks of participation. They should also be provided fairly to participants, with equal compensation for equal participation.

> **Vulnerable Populations**
> Any group unable to fully safeguard its members' self-interests in the context of a research project (e.g., minors, seniors, prisoners, ethnocultural minorities, persons with disabilities, those who are economically or socially disadvantaged, etc.)

Another ethical consideration for evaluators is whether they will be working with members from vulnerable populations. The Tri-Council policy states that "vulnerability is often caused by limited decision-making capacity, or limited access to social goods, such as rights, opportunities and power." Additionally, this policy notes that "Individuals or groups whose circumstances may make them vulnerable in the context of research should not be inappropriately included or automatically excluded from participation in research on the basis of their circumstances." Many access and retention programs specifically target vulnerable populations to increase their participation and success in PSE. When evaluating programs that serve members from such populations, it is especially important for researchers to ensure that their welfare is respected, and that they are being treated equitably.

> **Case Consideration: Success Beyond Limits' participatory approach**
> As previously mentioned, Success Beyond Limits (SBL) is performing Youth Participatory Action Research to evaluate their programming for youth in the Jane and Finch area in Toronto. By taking a participatory approach, members from the population being evaluated will be tasked with leading the evaluation. This approach will allow SBL alumni to assist in designing and implementing the evaluation so that they can ultimately understand the role of the SBL program on PSE trajectories and outcomes. Such a research design addresses the role of the researcher/evaluator when studying vulnerable populations.

Ultimately, when designing an evaluation that involves human participants, evaluators are advised to schedule ample buffer time so that they can plan for an ethical design, and apply for ethical approval accordingly. It is also recommended that this buffer time is factored directly into the project timeline, as the turnaround for receiving REB approval can often take a minimum of several weeks or months. Fortunately, patient evaluators will be rewarded with the opportunity for collecting rich information from evaluation participants.

---

25 For information on creating a consent form for potential participants, see the National Research Council Canada's underline.

Part 11: Building a Budget

Developing a budget is essential during the planning phase of an evaluation, as each budget will be highly customized to one's specific context. Evaluators might first consult human resources or other administrators to understand rules/regulations for constructing a budget within their institution or organization. They may want to consider the following questions:

- ✓ What funding is available for the evaluation?
- ✓ Do additional funds need to be acquired (e.g., fundraising or outreach)?
- ✓ What in-kind supports will be provided, and by whom? (e.g., staffing, meeting spaces, data storage, materials, equipment, etc.)

Upon establishing a general idea of the funds and in-kind supports available, evaluators should determine any non-negotiable elements of the budget, without which the evaluation could not answer the research question(s). Ultimately, it is imperative that evaluators understand how the budget might affect the study design, data collection, deliverables and timelines, and vice versa.

See Table 4 below for a sample budget template used for HEQCO projects. A typical project budget might include costs (real or estimated) for the following elements:

- ✓ Staffing (e.g., internal/external evaluators, researchers, analysts, other stakeholders, etc.)
- ✓ Travelling required for evaluation activities (e.g., attending meetings, collecting data, etc.)
- ✓ Accessing or storing data
- ✓ Materials, supplies and equipment
- ✓ Incentives for participants
- ✓ Knowledge dissemination (e.g., presentations, conferences, workshops, etc.)

Staffing is usually the most expensive consideration for an evaluation. To determine the pay rate per staff member, there are a few requirements: (1) an understanding of each staff's qualifications and expertise, (2) a rough estimate of the time that each staff will dedicate to performing evaluation activities, and (3) a good estimate of the length of the project as a whole. At the outset, evaluators should also establish who is responsible for overseeing the budget and administering funds, and whether there are any additional costs associated with this responsibility. Ideally, if payment(s) are directly associated with specific deliverables to be completed by staff (e.g., interim or final reports), they should also establish who will be responsible for approving these deliverables, how long will it take, as well as what will happen to the payment(s) if the associated deliverables do not meet the agreed-upon standards.[26] Estimating travel costs for evaluation activities will follow a similar process, and evaluators may consult human resources/administrators to determine the amount that can be reimbursed for travel expenses (e.g., public transit costs, flights, hotels and meal allowances).

When estimating costs for accessing or storing data, there will be some cases that are relatively straightforward, and other cases where evaluators might refer to data-sharing agreements. These agreements outline what data is being exchanged for a given project, alongside the associated costs, permissions for use, storage requirements and sharing privileges. Establishing costs for materials, supplies, equipment and incentives should also be relatively straightforward. But budgeting for knowledge dissemination can be tricky, especially if the results of an evaluation are not predictable. Evaluators might do well to hypothesize what events or publications would be warranted to attend/produce as a result of evaluation findings.

---

26 Evaluators should prepare for any potential delays in administering/receiving funds, and build these into the timeline accordingly.

Table 4: HEQCO's Budget Template

PERSONNEL:

Name of Proposed Team Members (insert additional rows if necessary)	Title/Position	# of Days/ Semesters	Per Day/ Semester Rate	Total: # x Rate
		Personnel Costs Subtotal		HEQCO: / In-kind:

DIRECT COSTS:

Detailed Description of Direct Costs (if applicable) (insert additional rows if necessary)	Cost
Direct Costs Subtotal	HEQCO: / In-kind:

TRAVEL AND ACCOMMODATION:

Detailed Description of Travel and Accommodation (if applicable) (insert additional rows if necessary)	Cost
Travel and Accommodation Subtotal	HEQCO: / In-kind:

OVERHEAD:

A detailed description of what constitutes overhead must be provided (if applicable) (insert additional rows if necessary)	Cost	10% of Cost
Overhead Subtotal		HEQCO: / In-kind:

FUNDING:

Personnel Costs Subtotal	
Direct Costs Subtotal (if applicable)	
Travel and Accommodation Subtotal (if applicable)	
10% of Overhead Total (if applicable)	
Firm Fixed Price (HEQCO Funding)	
HST (13%) (if applicable)	
Total HEQCO Funding (Firm Fixed Price + HST)	
In-Kind Contribution	
Total Project Cost (Total HEQCO Funding + In-Kind)	

Conclusion

Our aim for this manual was to present an overview of practical considerations for evaluating different types of programs. Program evaluation is both an art and a science. As discussed, the approaches, methods, tools and techniques must align with the research questions, in harmony with the larger vision and purpose for the evaluation. This alignment requires a great deal of careful consideration and judgement and we strongly believe that evaluation is worth such effort.

At the outset of this manual we identified the importance of program evaluation for HEQCO: Consistent, high-quality program evaluation is the only way we can confidently answer "what works" and "why does it work" in publicly-funded access and retention programs. We hope that this manual serves as a useful guide for developing evaluation plans for programs that seek to enhance access or retention in higher education. We look forward to working with proponents of HEQCO's Access and Retention Consortium (ARC) as they engage with the contents of this manual to develop, enhance or improve their program evaluations. We want to thank ARC proponents in advance for their interest in our shared mission of improving access, retention and success for students in Ontario's post-secondary education system.

References

Baker, A. (n.d). *Evaluation Services*: *Making Data Collection Decisions*.
http://nebula.wsimg.com/79cc3a812b2de18f3dad8393b80b0ea0?AccessKeyId=8A32C507409C3E494DCC&disposition=0&alloworigin=1

Better Evaluation (n.d.). *Evaluation Options: Collect and/or Retrieve Data.*
https://www.betterevaluation.org/en/rainbow_framework/describe/collect_retrieve_data

Better Evaluation (n.d.). *Evaluation Options: Develop Programme Theory / Theory of Change.*
https://www.betterevaluation.org/en/rainbow_framework/define/develop_programme_theory

Better Evaluation (n.d.). *Evaluation Options: Sample.*
https://www.betterevaluation.org/en/rainbow_framework/describe/sample

Better Evaluation (n.d.). *Evaluation Options: SWOT Analysis.*
https://www.betterevaluation.org/evaluation-options/swotanalysis

Canadian Evaluation Society (n.d.). *What is Evaluation?* https://evaluationcanada.ca/what-is-evaluation

Chen, H. T. (2014). Practical Program Evaluation: Theory-driven Evaluation and the Integrated Evaluation Perspective. United States: Sage.

Cozby, P. C., & Rawn, C. D. (2016). *Methods in Behavioural Research.* Canada: McGraw Hill Education.

Government of Canada (2014). *Tri-Council Policy Statement: Ethical Conduct for Researching Involving Humans.*
http://www.pre.ethics.gc.ca/pdf/eng/tcps2-2014/TCPS_2_FINAL_Web.pdf

Government of Canada (n.d.). *Supporting Effective Evaluations: A Guide to Developing Performance Measurement Strategies.* https://www.canada.ca/en/treasury-board-secretariat/services/audit-evaluation/centre-excellence-evaluation/guide-developing-performance-measurement-strategies.html

Government of Canada: National Research Council Canada (2012). *Guidance on Completing the Application for Ethics Approval.*
https://www.nrc-cnrc.gc.ca/eng/about/ethics_integrity/application_ethics_approval.html

Government of Canada: National Research Council Canada (2012). *Guidance on Completing the Consent Form Template.* https://www.nrc-cnrc.gc.ca/eng/about/ethics_integrity/consent_form_template.html

Innovation Network (n.d.). *Evaluation Plan Workbook.*
https://www.innonet.org/media/evaluation_plan_workbook.pdf

National Science Foundation (2002). *User-Friendly Handbook for Project Evaluation: An Overview of Quantitative and Qualitative Data Collection and Methods.* https://www.nsf.gov/pubs/2002/nsf02057/nsf02057_4.pdf

Project Management Docs. (n.d.). *Work Breakdown Structure.*
https://www.projectmanagementdocs.com/template/project-planning/work-breakdown-structure/#axzz5Mx5Sj1Ht

Project Manager (n.d.). *The Ultimate Guide to: Gantt Charts*. https://www.projectmanager.com/gantt-chart

Taylor-Powell, E. (1988). *Program Development and Evaluation: Sampling.* University of Wisconsin Extension.
https://learningstore.uwex.edu/Assets/pdfs/G3658-03.pdf

The Pell Institute (n.d.). *Evaluation Toolkit: Analyzing Qualitative Data*.
http://toolkit.pellinstitute.org/evaluation-guide/analyze/analyze-qualitative-data/

The Pell Institute (n.d.). *Evaluation Toolkit: Analyzing Quantitative Data*.
http://toolkit.pellinstitute.org/evaluation-guide/analyze/analyze-quantitative-data/

The Pell Institute (n.d.). *Evaluation Toolkit: Choose an Evaluation Design*.
http://toolkit.pellinstitute.org/evaluation-guide/plan-budget/choose-an-evaluation-design/

The Pell Institute (n.d.). *Evaluation Toolkit: Understanding Sampling*. http://toolkit.pellinstitute.org/evaluation-guide/collect-data/understand-sampling/

The University of Kansas' Community Toolbox (n.d.). *Operations in Evaluating Community Interventions: Choosing Questions and Planning the Evaluation.* Evaluating Community Programs and Initiatives. https://ctb.ku.edu/en/table-of-contents/evaluate/evaluate-community-interventions/choose-evaluation-questioins/main

Wingate, L., & Schroeter, D. (2007). *Evaluation Questions Checklist for Program Evaluation.* https://wmich.edu/sites/default/files/attachments/u350/2018/eval-questions-wingate%26schroeter.pdf

Additional Evaluation Resources

Balbach, E.D. (1999). Using Case Studies to Do Program Evaluation. Tufts University. California Department of Health Services. http://case.edu/affil/healthpromotion/ProgramEvaluation.pdf

Better Evaluation (n.d.). *Evaluation Options: Rainbow Framework.* https://www.betterevaluation.org/en/rainbow_framework

Ontario Centre of Excellence for Child and Youth Mental Health (n.d.). *Resource Hub.* http://www.excellenceforchildandyouth.ca/resource-hub

Stufflebeam, D.L. (2004). *Evaluation Design Checklist*. Western Michigan University Evaluation Center. https://www.wmich.edu/sites/default/files/attachments/u350/2014/evaldesign.pdf

The Pell Institute (n.d.). Evaluation Toolkit. http://toolkit.pellinstitute.org/

The University of Kansas' Community Toolbox (n.d.). *Evaluating Community Programs and Initiatives.* https://ctb.ku.edu/en/evaluating-community-programs-and-initiatives

U.S. Department of Health and Human Services Centers for Disease Control and Prevention: Office of the Director, Office of Strategy and Innovation. (2011). *Introduction to Program Evaluation for Public Health Programs: A Self-study Guide.* https://www.cdc.gov/eval/guide/cdcevalmanual.pdf

YouthREX (n.d). Evaluation Toolkit for Youth Programs. http://www.youthrex.com/toolkit/

Evaluation Groups

- Evaluation Capacity Network: http://www.evaluationcapacitynetwork.com/
- Evaluation Roundtable: http://www.evaluationroundtable.org/