



T O W A R D S T H E
PUBLIC'S HEALTH

Methods in Health Research and Evaluation

BY KIFFER G. CARD

About the Author



Kiffer G. Card, Ph.D., is an expert in public health methodology. His scholarly research focuses on the social, structural, and systemic influences shaping health and behaviour with a focus on policy change, health promotion, and harm reduction strategies for substance use, HIV, and Hepatitis C. Dr. Card's public health advocacy work focuses on the experience of individuals who are members of systemically oppressed and culturally devalued groups. He teaches courses in policy and program evaluation, quantitative methodology, and those relating to marginalized populations.

Suggested Citation

KG. Card. (2019). "Towards the Public's Health: Methods in Health Research and Evaluation." 1st Edition. HLTH 506: Health Research and Evaluation, University of Victoria, Victoria, British Columbia, Canada.

Table of Contents

Chapter 1.	The Necessity of Public Health Research and Evaluation	1
Chapter 2.	Preliminary Evaluations	30
Chapter 3.	Process Evaluations	73
Chapter 4.	Outcome Evaluations	87
Chapter 5.	Quantitative Data Collection	110
Chapter 6.	Quantitative Data Analysis	133
Chapter 7.	Qualitative Data Collection	143
Chapter 8.	Qualitative Data Analysis	152
Chapter 9.	Mixed Methods Designs	161
Chapter 10.	Recruiting Participants	165

Chapter 1

THE NECESSITY OF PUBLIC HEALTH RESEARCH AND EVALUATION

Learning Objectives

- Discuss the necessity of evidence-informed practice and evaluation.
- Differentiate between research and evaluation in Public Health.
- Identify steps of a policy or program evaluation.

JOBS ^{not} JAILS

Part 1

Why conduct research or evaluations?



Source: <https://www.religiousleftlaw.com/2012/04/cultivating-hope-compassion-at-homeboy-industries-.html>

Evidence-based Practice

This section is adapted from *Freakonomics's* podcast, "When Helping Hurts", Welsh (2017), & Vosburgh (1980)

Consider the following statement by Gregory Boyle, the founder and Director of an organization called *Homeboy Industries*:

"If you're driven by outcomes you're going to only work with a population that will give you good ones. We're going to work with the most belligerent, difficult, hard-headed, as long as they walk through the door."¹

Homeboy Industries is a youth program founded in 1992 that assists high-risk youth, former gang members, and recently incarcerated individuals in accessing mental health counselling, legal services, education, and employment. If you go to the website of Homeboy Industries, you'll read a slightly different and more eloquently stated version of the quote from above:

“What if we ceased to pledge our allegiance to the bottom line and stood, instead, with those who line the bottom.”²

In either of these statements, we are confronted with what seems like a natural dichotomy: the bottom line (*dollars, numbers, statistics, and outcomes*) versus communities in need (marginalized and oppressed individuals, families, and neighborhoods). But what if this contradiction were an illusion – a type of propaganda, so to speak? What if, in fact, the best thing for the bottom line really is the best thing for communities in need?

You may very well be unconvinced that this is the case, but if we start with the assumption that a more efficient and effective public health care system is better for everyone, then the next question we have to ask ourselves is “How do we make the healthcare system more effective?”

This is an important question. For example, if you ask Gregory Boyle whether Homeboy industries works, he will tell you some formulation of the following:

“no gang member would walk through the door if this place didn’t help, if it didn’t somehow work!”¹

This may be true. However, the existence of a program – and even the successful implementation of a program – is not necessarily sufficient evidence that a program meets its goals and objectives. When a program has proven its ability to meet its goals and objectives it can be said to be “evidence-based.” Indeed, to understand if a program works, you have to have some standard or outcome against which it can be evaluated. With that in mind, the focus of this text is to discuss how such evaluations can be used to make the healthcare system work for all those who are vested in its success.

Case Study: The Cambridge-Somerville Youth Study

This section is adapted from [Freakonomics’s podcast, “When Helping Hurts”](#), Welsh (2017), & Vosburgh (1980)

One of the earliest evaluations undertaken of a program begins with Richard Clarke Cabot. Cabot was a Harvard physician and pioneer in Social Work. Among the many varied accomplishments Cabot is known for was his commission of the *Cambridge-Somerville Youth Study (CSYS)* in the 1940’s – during the great depression.

Unlike most social programs at the time Cabot's study a longitudinal survey. Today, the survey continues to follow some of the original participants in the CSYS. Why, you might ask is a study started in the 1940's still going on? To understand the answer to this question you have to go back to Boston in the late 1930's. Juvenile delinquency, incarceration, and recidivism rates among youth were staggering.³ Cabot's idea was to use a strategy called "directed friendship" as a way to provide kids with mentors, who would hopefully help keep them out of trouble by meeting with them every couple of weeks, helping them with their homework, and taking them to the YMCA. In the summer, youth in the treatment group would also go to a summer camp where they could get away from the city. But, Cabot wanted to really know whether this would be the case: Would providing youth with simple mentorship and a summer camp actually help them?

To answer this question, Cabot incorporated a longitudinal experiment as part of his program. It began with 250 youth who were identified by their preachers, teachers, and parents as being a bit troubled or prone to getting into trouble and 250 youth who seemed to be doing well. He then matched these kids based on characteristics such as family size, neighbourhood, and income and he randomly assigned one of the youth to participate in the study and the other to a control group. The youth in the control group received no mentorship or services while the one in the treatment group participated in the mentorship program.

To clarify, kids who were doing well were matched to kids who were not doing well and then the pairs were split up randomly into either the control group or the treatment group. Therefore, the treatment group had some kids doing well and some kids who were struggling, as did the control group.

After six years, Cabot had a huge number of records that assessed a variety of life outcome measures, such as alcoholism, life satisfaction, and incarceration. However, in a milestone 1948 analysis, conducted when the boys were in their 20's, Cabot found no significant differences in the outcomes of interest for those who participated in the study compared too those who did not!

This makes you wonder: was the mentoring program a waste of time? After all, the money spent to help these youths could have been spent on other programs – maybe even paying for tuition for these young men so they would have the opportunity to get ahead. This is called an opportunity cost, which can be more

broadly defined as the foregone value and gains that are lost when one alternative is selected over another.

Opportunity costs aside, the 1948 results are not the end of the Cambridge-Summerville story. In the 1950's Joan McCord, the first woman to become President of the American Society of Criminology, took lead on the study. At this time, most men were in their 40's and had been lost to follow-up, but McCord took it upon her self to reconnect with the original participants – eventually contacting 98% of them. So what did McCord find? Well, when asked, participants described the program as life changing and celebrated its positive and formative influence. Yet, when measured against mortality measures, criminal histories, job satisfaction, and marital satisfaction, the treatment group was significantly worse off compared to the control group. In other words, participation in the CSYS led to worse outcomes for the men who had participated as youth. In fact, there was a significant dose-response effect whereby those who were in the treatment longer were worse off than those who were in for shorter periods of time.

While the exact causes for the CSYS's negative impact on youth is hard too parse out, one thing is made clear in thinking about CSYS's legacy: public health and social programs absolutely need to be evaluated – not only to make sure they are economically effective, but to ensure they are not doing hidden harm to those who participate.

Learning Activities

1. Imagine you are leading your health authority's implementation of *Home First*, which is a policy increasingly used in jurisdictions to free up hospital beds by helping frail patients get out of hospital and back into their homes as soon as possible. It's meant for patients who have completed acute care treatment and no longer need 24-hour attention in hospital, patients who can heal safely at home with the right kinds of support. Home First connects you to that support with enhanced home care services for up to 90 days. What are some reasons you would want to conduct an evaluation of the Home First policy in your jurisdiction?
2. Pre-exposure prophylaxis (PrEP) for HIV is a prevention strategy that can prevent HIV acquisition in HIV-negative persons who take antiretroviral therapy drugs. In addition to being effective, Tenofovir-Emtricitabine, the combination antiretroviral drugs used in PrEP have been shown to be fairly well tolerated and safe. There is near universal scientific consensus that PrEP should be implemented among populations at high risk for HIV. In 2018, British Columbia began coverage for PrEP for individuals identified by their physician at high risk. Given the scientific support for PrEP, should money be allocated to conduct an evaluation of British Columbia's new policy? Why or Why not?
3. Review the additional readings assigned to this chapter and write two or three paragraphs reflecting on the role of evaluation in contemporary public health policy-making and programming. In your reflection be sure to engage actively with each reading.
4. Would you support a government policy that restricted grant and programmatic funding to studies that had an evaluation component that provided clear indicators against which the success or failure of a program could be measured? Why or why not?

Additional Readings

- Smith & Petticrew. (2010). "Public health evaluation in the twenty-first century: time to see the wood as well as the trees." *Journal of Public Health*.
- Wanless. (2004). "Securing Good Health for the Whole Population." United Kingdom Treasury.
- Spiegelman. (2016). "Evaluating Public Health Interventions: 1. Examples, Definitions, and a Personal Note." *American Journal of Public Health*.
- Connell et al. (2015) "'Dark logic': theorising the harmful consequences of public health interventions." *Journal of Epidemiology and Community Health*.



Part 2

What is evaluation?

Source: <https://medicalxpress.com/news/2019-01-uk-female-surgeons-workplace-discrimination.html>

Public Health Research: Epidemiology and Evaluation

Recognizing the importance of promoting evidence-informed practices in the health sector, public health leaders across Canada have paid increasing attention to how public health programs and policies are implemented and evaluated. They have adapted traditional research strategies and applied them to the scientific studies of programs.

While this push is not new, many continue to think of public health research in terms of the classical epidemiological studies that examine the distribution and determinants of disease within a population. You can probably readily recall examples of epidemiological research introduced in your introductory public health courses: John Snow's Cholera investigation or the Framingham Heart Study's famous research on cardiovascular disease.

Epidemiological studies generally use cross-sectional, case-control, cohort studies, or some variation on these traditional designs to understand the trends and patterns underlying health related states and events. However, public health research on implementation and evaluation is not necessarily well-aligned with

these traditional epidemiological studies. Yet they are a fundamental, if not even more important, feature of contemporary public health work.

With this in mind, I want to help us distinguish between research and evaluation. Of course, I should also recognize that the similarities and differences between research and evaluation have been hotly contested and there are many conflicting points of view among practitioners of research and evaluation. According to the Canadian Institutes of Health Research,

“Proponents of [the position that research and evaluation are different] cite such factors as the centrality of ‘valuing’ to evaluation; the inherently political nature of evaluation activities; the limited domain of application of evaluation findings (local and specific rather than transferable or generalizable); and the important role of theory in research compared to evaluation activities.”

In this text, I generally take the approach that research and evaluation are distinct practices and disciplines. However, the emergence of *implementation science* has called into question this dichotomy as there is now a growing focus, even among researchers, on the localized processes that make evidence-based practices work.

If you are feeling a bit unsure of the difference between epidemiology and program evaluation, let me give you a couple examples (which may or may not clarify things for you):

- **An epidemiological research study** might compare the prevalence of depression symptomology in men and women and report that women are more than twice as likely to have depression, as measured by doctor-diagnosed cases of depression.⁴ This study is an epidemiological study because it is concerned with the distribution of disease in a population. Such a study would likely follow the traditional scientific method: it would start with asking a question, constructing a hypothesis and then move through to testing a hypothesis and analyzing the results of that test. The hopes of such a study would probably be to generalize the results to other settings or to make an assessment of the underlying biological or social phenomena at work.

- **An evaluation**, on the other hand, might examine whether there are gender differences in how antidepressants are prescribed following the implementation of a continuing education course that intends to eliminate gender-based differences⁵ that contribute to over diagnosis of mental health conditions in woman and underdiagnoses in men. In this case, the study is less concerned about the distribution or determinants of disease and instead focuses on whether the delivered intervention (i.e., the continuing education course) is meeting its goal. Therefore, it is an evaluation study. It focuses on the value of a localized phenomena. The intent is not to generalize a result to other settings, but to uncover and ascertain an honest assessment of the continuing education course.

You can see that in both cases, the primary health outcome of interest was depression and the primary explanatory variable of interest was gender. However, the first was concerned with what characteristics were associated with depression, and the second was concerned with whether a public health policy (i.e., the policy to prescribe based on symptomology and not gender) or program (i.e., the continuing education course teaching doctors how to overcome gender bias in diagnosing men and women) were effective.

For the most part, I will refer to evaluation throughout this text, with the implicit recognition that, in many circles, evaluation is simply a practice-oriented form of research. However, the focus of this text is strongly oriented towards the concept of evaluation as a methodology aimed at assessing value. For our purposes, value can be conceptualized as three interrelated issues: merit (i.e., quality), worth (i.e., cost-effectiveness), and significance (i.e., importance).

Fundamentally, implementation and evaluation studies focus on improving health services, while epidemiological studies focus on measuring trends, patterns, and occurrence of health related states or events.

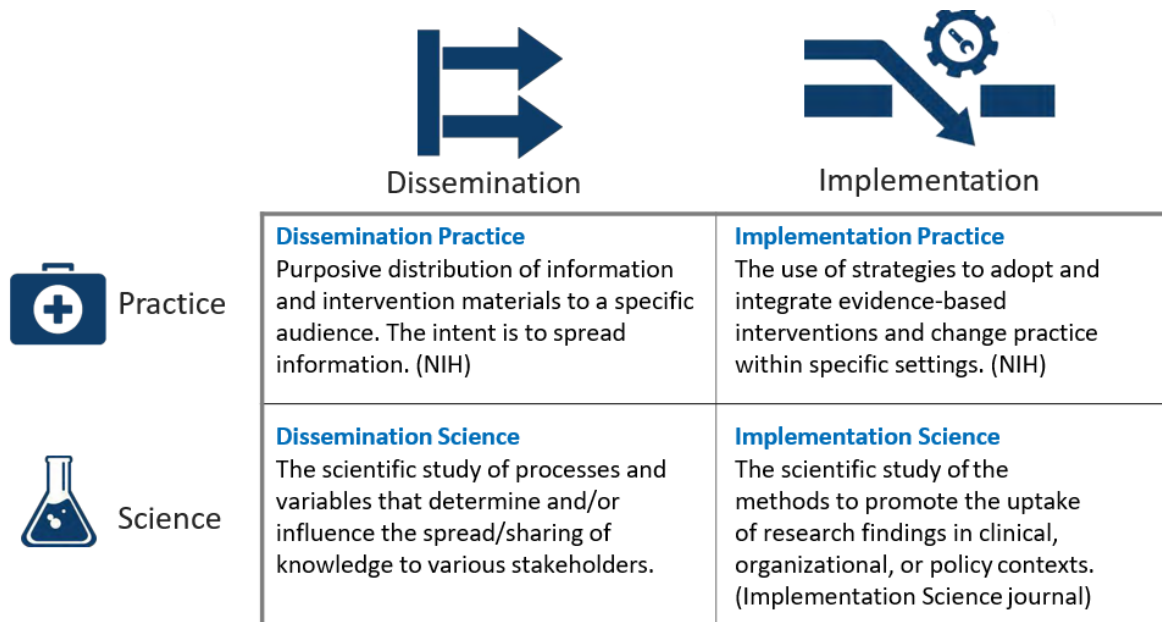
Considering the distinction, I've drawn here, it might be tempting too think of evaluation as a "small scale" version of research. However, I would warn that evaluations can occur across all levels of the social ecology – from project and program level evaluations, to organizational evaluations, community level evaluations, and even evaluations at the local, state, or national level. Furthermore,

evaluations can be focused on specific programs or on broad policies that impact millions.

The Emergence of Implementation Science

As mentioned above, the emergence of program science, or implementation science, has pushed back against the dichotomy between research and evaluation. One model, shown in **Figure 1**, distinguishes between practice and science as well as implementation and dissemination. In this framework, *dissemination* is defined as the distribution of an intervention, sometimes through the creation of manuals, presentations, or websites while *Implementation* involves the integration of a new practice into a new setting. Each of these can then be broken down into *practice*-based activities and *scientific* activities. Practice-based activities are focused on the things organizations are doing whereas scientific activities are focused on understanding the process and developing a theory for why it works or does not work.

Figure 1. Implementation and Dissemination Practice and Science



So *dissemination practice activities* might include developing and distributing policy briefs and research summaries to policy makers or the collaboration between a profession association and provincial agencies to make sure healthcare professionals represented by the association had received copies of their new

clinical guidelines. Meanwhile, an example *dissemination science activity* might include comparing two different website layouts to determine which format patients find more useable. On the flip side, implementation science activities might include a group wants to evaluate the most effective strategies to train new staff to effectively use the screening tools related to their role or a researcher assessing which tool helps counselors have better adherence to cognitive behavioural therapy. Finally, an implementation practice activity might be exemplified by health facilities using a set of tools, training, and support to decrease staff practices that were associated with hospital acquired infections.

With these sorts of classification schemes in mind, it is important to become familiar with the lingo used in your specific settings. However, I want to stress that you should not get too caught up in these sorts of distinctions. While knowing the terminology will help you to be able to find resources specific to whatever you are trying to do, it is ultimately up to you to arbitrate what specific activities you include as part of your evaluation and research. As I noted above, this text recognizes the distinction between evaluation and research. However, we will focus mostly on evaluation activities – as they tend to be a bit more applied and a bit more relevant to the types of public health careers that many MPH students go on to pursue. Yet, you should recognize that the overlap in what is actually done in a research study will be in many ways very similar to what you would do as an evaluator. Furthermore, I personally believe that the distinctions are disintegrating as researchers adopt a more community-focused approach to their work.

Major Types of Policy and Program Evaluation

You are probably not too surprised to find out that public health evaluations can be further broken down into multiple types of evaluation – each with its own role in the evaluation process. Understanding the different types of evaluations used in public health will not only help you to understand the terminology used in describing public health research but will also illuminate some of the primary issues that public health practitioners are concerned with when putting evidence into action. It is my hope that promoting a greater understanding of these different types of evaluation will help our sector move *towards the public's health* by making the health system more effective at meeting the needs of all those who interact with it.

While evaluation strategies have been categorized into many classification schemes, in public health, the most widely used classification scheme distinguishes between four types of evaluations (1) the preliminary evaluation, (2) the process or implementation evaluation, (3) the outcome or effectiveness evaluation, and (4) the impact evaluation. With terminology coined by Michael Scriven in 1967, these four categories are sometimes further classified into either “*formative*” (i.e., those taking place as the program is being *formed* or implemented) or “*summative*” (i.e., those taking place *summarizing* the results of a program) evaluations

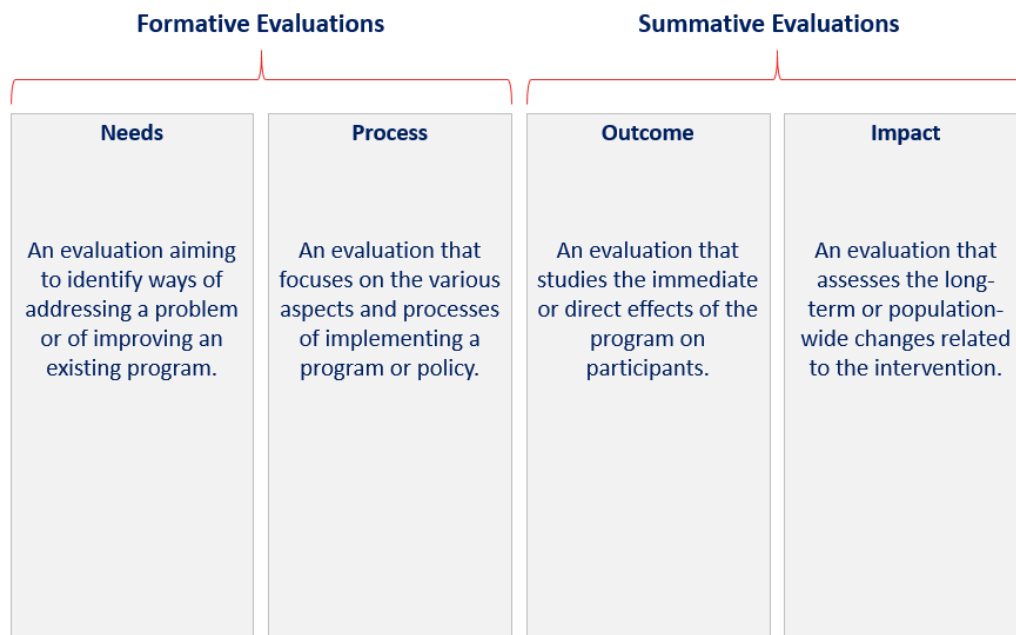


Figure 1. Broad types of Evaluations conducted in Public Health Programs

Preliminary Evaluations

Preliminary evaluations, also called formative evaluations or needs evaluations, evaluate the urgency of an intervention as well as concepts such as **acceptability** (will those who participate in or operate an intervention be willing to), **appropriateness** (will the intervention meet its objective), and **feasibility** (will the intervention be implementable) for the various programs or policies that might be implemented. The primary goal of a preliminary evaluation is to ensure that a new or adapted program or policy is correctly selected and ready to succeed in a given environment. Preliminary evaluations also include “Needs Assessments” and can help you when it is not fully understood which intervention or features of an

intervention should be adopted. We will discuss preliminary evaluation strategies later in this text.

What it shows:	Why it is useful:	When to use it:
<ul style="list-style-type: none"> • Needs • How target audience receives information • Ideal champions/ spokespeople • Challenges in using program content 	<ul style="list-style-type: none"> • Making sure there's a need • Clarification • Make revisions • Maximizes the likelihood that the program will succeed 	<ul style="list-style-type: none"> • Program development or revisions • Examples: <ul style="list-style-type: none"> • Needs assessment • Usability testing • Implementation planning • Stakeholder engagement

Process Evaluations

Once a program begins to launch, a process evaluation can (and should) begin. The purpose of the process evaluation is to study how well the program or policy is being implemented. Not only does this tell you whether a study is being implemented as planned (a concept known as **fidelity**), but it can act as a sort of early warning system if something is not working well. The process evaluation may include some interviews patients, but largely, these evaluations focus on patient-experience and not whether the intervention being evaluated is having the desired effect on patients. We will discuss process evaluation strategies in greater detail later in this text.

What does it show?	Why is it useful?	When is it done?
<ul style="list-style-type: none"> • Activities conducted • Characteristics of activities conducted <ul style="list-style-type: none"> • (e.g., how many people are participating in the program and how many people are not) 	<ul style="list-style-type: none"> • Identifies challenges in reaching the target population • Allows programs to evaluate how well their plans, procedures, activities, and materials are working and to take timely, corrective actions 	<ul style="list-style-type: none"> • As soon as the program begins

Outcome Evaluations

Outcome evaluations usually take place after patients have completed the intervention. Outcome evaluations focus on the most immediate results of a project while impact evaluations focus on whether the program achieved its ultimate goal. Often times, outcome evaluations distinguish between **outputs** and **outcomes**. In this typology, outputs are the results achieved directly as a result of the intervention. For example, an output might be the number of people who were educated about the importance of getting screened for sexually transmitted infections following the introduction of sexual health counseling in a community clinic.

Outcomes, on the other hand, focus on whether the program is meeting its goal. In our example, an outcome might be something like more frequent STI testing among those who received sexual health counselling from their doctor or nurse. In summary, outputs are reached immediately as a program or policy is implemented, and outcomes are the mid-term results so to speak. We will discuss outcome and impact evaluations later in this text.

What it shows:	Why it is useful:	When to use it:
<ul style="list-style-type: none">• The degree to which a program is meeting its goals• For example, changes in knowledge, skills, attitudes, beliefs, behaviour, patient/other outcomes	<ul style="list-style-type: none">• To make adaptations• Is program moving toward achieving goals	<ul style="list-style-type: none">• After the program has made contact with at least one person or one group of people in the target population

Impact Evaluations

In addition to outputs and outcomes there are also **impacts**. Like outcome evaluations, Impact evaluations tend to take place after a study has completed. Impact evaluations tend to be more long term or larger scale as they focus on the overall effect of a program in addressing the issue under consideration. Once again continuing our example from above: the impact for an STI screening program might be a reduction in the incidence of syphilis at the population level that is attributable to the intervention being studied.

As you have likely realized at this point, the differences between these evaluations are closely linked to four broad stages of a program or policy’s life cycle. Preliminary evaluations usually occur first, followed by process evaluations, then outcome evaluations, and, finally, impact evaluations. However, the life cycle of an evaluation is not the only useful way to think about evaluations. We must also consider the values against which policies and programs are judged and how these values are incorporated into the evaluation itself.

What does it show?	Why is it useful?	When is it done?
<ul style="list-style-type: none"> • Activities conducted • Characteristics of activities conducted • (e.g., how many people are participating in the program and how many people are not) 	<ul style="list-style-type: none"> • Identifies challenges in reaching the target population • Allows programs to evaluate how well their plans, procedures, activities, and materials are working and to take timely, corrective actions 	<ul style="list-style-type: none"> • As soon as the program begins

The Knowledge-to-Action Cycle

In the Canadian context, the Knowledge-to-Action cycle (**Figure 2**) is a widely used framework for guiding evaluation and research activities. The types of evaluations discussed above can be mapped onto this framework quite easily. Formative evaluations aim to identify the problem; determine the know/do gap; identify, review, and select knowledge; adapt knowledge to local context; and assess barriers/facilitators to knowledge use. Process evaluations may repeat some of these activities, but also focus on selecting, tailoring, and implementing interventions; monitoring knowledge use; and sustaining the use of knowledge. Outcome and impact evaluations then tend to focus on evaluating outcomes and may replete some of the earlier activities, such as a re-assessment of the barriers and facilitators experienced by the research team.

If we were to try and enforce the research/evaluation dichotomy, research would focus on knowledge inquiry, synthesis, and the production of tools and products; whereas evaluation would tend to focus on the rest. Of course, this sort of rigid breakdown is increasingly irrelevant given the broad activities that researchers now engage in as part of their scientific research.

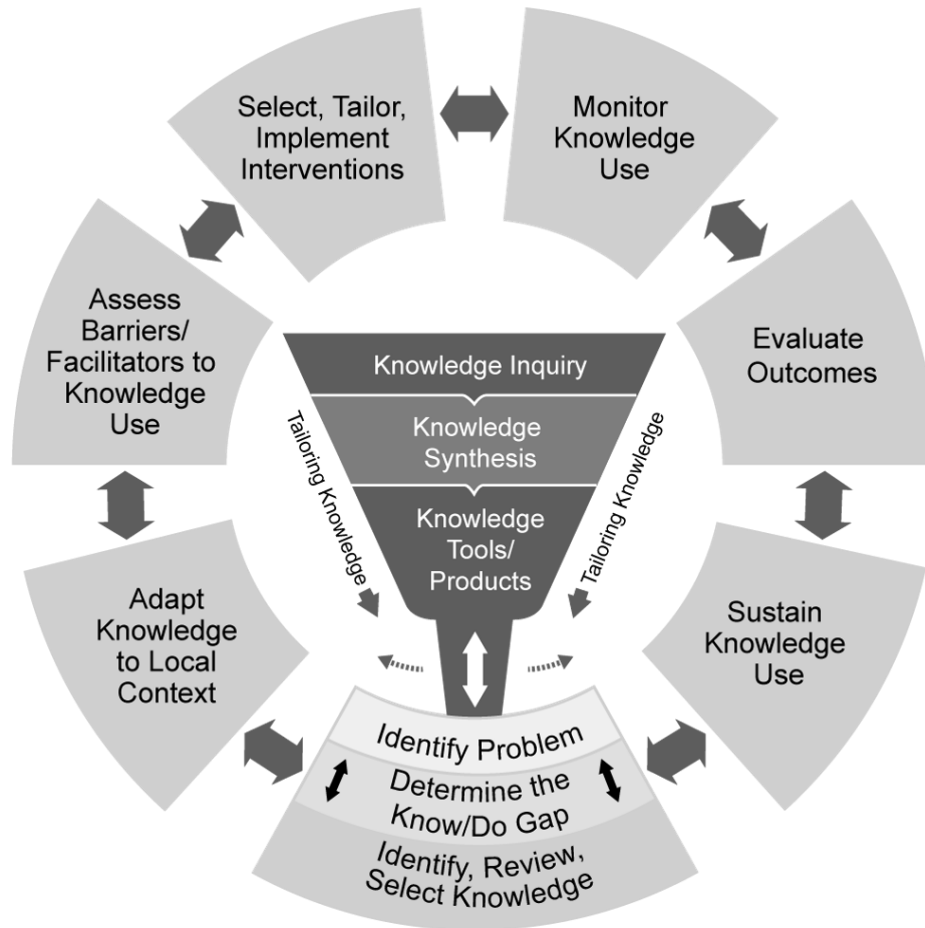


Figure 2. The Knowledge to Action Cycle

Values underlying an evaluation

The four types of evaluation discussed above reflect a strong and prevailing emphasis on outcomes in the evaluation process. However, evaluators and the organizations that sponsor them generally have broad leeway to determine the values that guide their evaluations. As such, it is useful to think through some of the values that might shape your future evaluation work. Below I have provided a brief list of some popular values that guide evaluations:

Goal and Objective-based Evaluations

Objectives are perhaps the most obvious and widely used value that underlies many public health evaluations. Later in this text we will talk more about goals and objectives, but suffice it to say here that many evaluations are conducted pragmatically with a very specific end in mind. These ends can be shaped by

economic or budgetary pressures, demands from granting agencies and funders, or the personal whims of the individual stakeholders with decision making responsibilities for the organization.

Criteria-based Evaluations

Related to the value on objectives, another frequently encountered value is criteria. Criteria can be selected by the evaluator or advanced by the stakeholders sponsoring the evaluation. Example criteria often include efficiency (i.e., cost, time), relevance/appropriateness, effectiveness, impact, and sustainability. Other criteria can be added, but these are generally criteria that you will need to consider in every criteria-based evaluation. It is the job of the evaluator to take criteria and identify indicators by which the criteria can be measured.

For now, it's important to recognize that criteria-based evaluations are useful when comparing several potential options for an intervention. Additionally, criteria-based evaluation increases the openness and transparency of an evaluation because it clarifies on what terms an evaluation is being conducted. Relatedly, these evaluations tend to be focused and provide a foundation for dialog with stakeholders. Furthermore, they help decision makers to systematically evaluate the results you present to them at the end of your evaluation.

Theory-based Evaluations

Occasionally, it is important for evaluators to adhere to a specific theory, model, or framework when they are conducting a review.⁶ Theory-based approaches are useful when limitations and constraints make it difficult to fully assess important criteria or when it is not entirely clear what should be measured. They address these challenges by following a theory of change, which helps to organize and guide the evaluation of a program. Widely-used frameworks are provided in **Table 1**.

For example, if you are conducting a social media campaign aimed at reducing sexual violence, you may or may not be able to directly measure some of the most fundamental aspects and processes that impact your intervention. You might be able to assess impact by conducting a population survey to assess the prevalence of sexual violence; and you are very likely to know the reach of your intervention.

Table 1. Common Theories, Models, and Frameworks for Evaluation

TMF	Description	Citation
Knowledge to Action (KTA)	A process model that defines and describes Knowledge Translation processes and outlines strategies for building Knowledge Translation capacity	Graham, D., et al. (2006). Lost in translation: Time for a map? Journal of continuing education in the health professions.
Capability, Opportunity, Motivation – Behaviour (COM-B)	A behavior change theory focusing on capabilities, opportunities, and motivations that proceed a behaviour.	Michie, S. et al. (2011). The behaviour change wheel: A new method for characterizing and designing behaviour change interventions. Implementation Science.
Theoretical Domains Framework (TDF)	A framework that was developed to make behavior change theories more accessible in relation to implementation practices.	Cane, J., et al. (2012). Validation of the theoretical domains framework for use in behaviour change and implementation research. Implementation Science.
Consolidated Framework for Implementation Research (CFIR)	A framework that can be used to understand the implementation context	Damschroder, L., et al. (2009). Fostering Implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Science.
Quality Implementation Framework (QIF)	A framework that provides a synthesized overview of critical steps of implementation that can be used as a guide for practice and research.	Wandersman, et al. (2012). The Quality Implementation Framework: A Synthesis of Critical Steps in the Implementation Process. American Journal of Community Psychology.
Interactive Systems Framework (ISF)	A framework that outlines the people/organizations/contexts needed to move effective prevention into the field.	Wandersman, A., et al. (2008). Bridging the gap between prevention research and practice: The interactive systems framework for dissemination and implementation. American journal of community psychology.
RE-AIM	A framework for evaluating public health interventions that assesses 5 dimensions: Reach, Efficacy, Adoption, Implementation, and Maintenance. These dimensions occur at multiple levels (e.g., individual, clinic or organization, community) and interact to determine the public health or population-based impact of a program or policy.	Glasgow RE, Vogt TM, Boles SM (1999). Evaluating the public health impact of health promotion interventions: the RE- AIM framework. Am J Public Health.

However, everything in between – the theory of change, social network dynamics, cultural shift, and so fourth – are very difficult to measure adequately. Therefore,

an evaluation of this program would have to provide a theory for the link between number of people reached and population prevalence of sexual violence.

Based on careful review of the evidence generated from a theory-based evaluation, an evaluator might refine the theory and provide recommendations for how the program can be revised to account for the observed evidence. Theory-based evaluations are often contrasted with so called “black box evaluations” in which the theory or mechanisms of causality that underlie the programs design are not considered (i.e. the focus is purely objectives based).

Goal Free Evaluations

In some settings, a goal free evaluation might be what is valued. These evaluations typically do not have predefined goals or objectives, but they may have a scope of inquiry that is limited to a particular aspect of the program. You might wonder why this would be an acceptable approach? Usually, these sorts of evaluations are useful when something is not working, but it is unclear what.

Take for instance an evaluation of workplace health and morale. The management team might recognize that there are serious morale issues in their ranks, but are unsure what is driving these concerns. They might bring in an evaluator and tell them that the scope of their evaluation is workplace health and morale and ask the evaluator to design an evaluation program that can ascertain some of the central issues and how they can be addressed. In this case, a goal free evaluation is useful because the management team does not want to bias the evaluator in their evaluation.

As an alternative scenario, the sponsors of a goal free evaluation might not want to disclose the goals of their program to the evaluator out of concern that knowledge of the goals will bias or shape the evaluation in some way. This neutral stance allows for a more independent evaluation of a program not bogged down by pre-determined expectations for how the program should operate. In many ways, this is the antithesis of the theory-based evaluation.

One of the obvious criticisms levied against goal free evaluations is the lack of established methodology. In truth, there are only two requirements for a goal-free evaluation. The first is that the evaluator is external and the second is a screener who acts as a mediator to remove goal-oriented communication between program

sponsors and the evaluation team. In this sense, the goal free evaluation is similar to a blinded study.

Participatory Evaluations

Increasingly, evaluations in public health are becoming participatory. Indeed, the role of stakeholders in public health evaluations has become more and more integral to contemporary evaluation protocols. This is out of recognition that stakeholders are not only necessary for actually conducting an evaluation (you will always need their help), but that they also tend to be the most reliable experts. Therefore, consulting them can help you ascertain the best insights that will help you make key recommendations necessary to improving the policy or program you are working on. We will discuss the role of stakeholders throughout this text. However, despite the emphasis of this text on involving stakeholders at the outset of the design, participatory evaluations engage a broad swath of interested and stakeholders throughout the entire evaluation process.

Expertise-oriented Evaluations

In closing this section on values, we should not omit the oldest, and perhaps a bit dated, value in public health evaluations: expertise. Expertise-oriented evaluations usually take form when an outside expert is invited to conduct an evaluation. This is frequently the case in government commissioned reports and evaluations. The rationale for this approach is that an expert in a field will bring considerable credibility to the evaluation. In these cases, the expert is asked to judge the quality of a program or policy. They are often given broad leeway to guide the evaluation, but will often be governed by terms of reference with limited scope and specified outputs. Frequently, the evaluator is usually external.

Evaluations may take on multiple aspects of any one these evaluation types. Indeed, this is why I have tried to frame them as values, more than discreet protocols for design. For example, a criteria-based evaluation very well may be participatory. I recommend that each of these evaluation values be considered in negotiating with the evaluation sponsor what the evaluation should look like.

Learning Activities

1. In your own words explain how epidemiological research differs from evaluation?
2. Match each of the research questions below with the type of evaluation it is most closely related (preliminary, process, outcome, impact). Explain why you chose the option you selected. If more than two options are possible, what characteristics added to the scenario would influence whether it fit definitively in one category over the other?
 - a. How many elderly patients in the intensive care unit have been screened for long-term care?
 - b. Has the average recovery time for hip replacement patients decreased after the introduction of a new physical therapy program?
 - c. Do people who participate in sex work need access to free condoms?
 - d. What was the five-year mortality rate ratio for breast cancer patients who participated in a social support program?
 - e. Which opioid agonist pharmaceutical drug had the lowest relapse rates?
3. Identify or develop a scenario for which a program evaluation would be required. After developing a scenario with sufficient detail, share your scenario with a partner and discuss the following questions:
 - a. What values would be appropriate for an evaluation addressing this scenario? Why?
 - b. What elements of the scenario would impact your decision to choose this value as a guiding principle of your evaluation?"
4. Compare and contrast the Case Studies presented by Youker et al. (2016) and Nelson & Schreiber (2009). What aspects did you like or dislike about each evaluation approach? Do you think that evaluation strategies presented in these two scenarios are interchangeable?

Additional Readings

- Wander (2019). "*What is evaluation? Perspectives of how evaluation differs (or not) from research.*" Psyarxiv Pre-print.
- National Centre for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. (2007). "Types of Evaluation." United States Centers for Disease Control and Prevention. Department of Health and Human Services.
- Youker et al. (2016). "Who Needs Goals? A Case Study of Goal-Free Evaluation." Journal of Multi-Disciplinary Evaluation.
- Nelson & Schreiber. (2009). "Participatory Evaluation: A Case Study of Involving Stakeholders in the Evaluation Process." Visitor Studies.

Part 3

How do you do an evaluation?



Source: <https://upswell.org/wp-content/uploads/2018/10/Focus-Group-10-15-18-1600x900.jpg>

Steps in a Public Health Evaluation

It is admittedly a bit absurd to try and provide a comprehensive and universal list of steps in a program or policy evaluation. There are probably as many protocols for evaluation as there have been evaluations – and, indeed, you as a public health researcher and evaluator will have to identify the steps in the evaluation process that work best for you and the stakeholders you are working with. What I hope to provide here is a framework of the general steps that will help you think through the evaluation process. These are meant to be helpful, not proscriptive. That said, missing one of these steps can spell doom for your evaluation by shaping how it is used by decision makers to inform their programming decision.

To begin, consider the following three lists of evaluation “steps” (**Table 1**). The left most list is provided by the U.S. Center for Disease Control, the middle list provided by the Canadian Institutes of Health Research, and the list on the right is provided by the WHO.

Table 1. Evaluation Frameworks by the U.S. CDC, CIHR, and WHO

U.S. CDC	CIHR	WHO
Step 1: Engage Stakeholders	Step 1: Consider the purpose of the evaluation	Step 1: Defining evaluation questions and criteria
Step 2: Describe the Program	Step 2: Identify Intended users of the evaluation	Step 2: Preparing terms of reference
Step 3: Focus Evaluation Design	Step 3: Create a structure and process for collaboration	Step 3: Choosing a methodological approach.
Step 4: Gather Credible Evidence	Step 4: Assess evaluation capacity and build an evaluation team	Step 4: Estimating resources
Step 5: Justify Conclusions	Step 5: Gather relevant evidence	Step 5: Determining the evaluation management structure
Step 6: Ensure Use and Share Lessons	Step 6: Build a shared level of consensus about the evaluations	Step 6: Managing conflicts of interest
	Step 7: Describe the program or intervention	Step 7: Establishing an evaluation work plan
	Step 8: Confirm purpose, focus the evaluation	Step 8: Preparing the inception report
	Step 9: Identify and prioritize evaluation questions	Step 9: Identifying information needs and data collection methods
	Step 10: Select methods and data sources	Step 10: Briefing and supporting the evaluation team
	Step 11: Clarify resources and responsibility areas	Step 11: Ensuring Quality
	Step 12: Implement evaluation activities	Step 12: Preparing the draft evaluation report
	Step 13: Communicate evaluation findings	Step 13: The final evaluation report.

The first thing you will notice is the CDC's list is more concise and many of the activities described by the CIHR or WHO are simply collapsed into a single step under the CDC's framework. You may have also noticed that the CDC list starts with engaging stakeholders, while the others highlight the leg work required to simply define what the evaluation aims to accomplish.

Ultimately, the exact order of these steps will depend on context and your position relative to the program (i.e., whether you are an internal or external evaluator) and how the evaluation came to be (e.g., commissioned, funder-initiated, investigator-initiated, regular cycle). If you find other frameworks you will notice that there is no definitive way to categorize these steps. Therefore, I would not get caught up in these sort of details if I were you. The important thing is that you are thorough enough to meet the evaluation standards desired by the organization or person sponsoring the evaluation.

Indeed, as you plan, implement, and utilize your evaluation results, the guiding framework you choose is often up to you. Under some circumstances, your organization may already have a framework for conducting evaluations that it uses across its programs. For example, Public Health Ontario provides [10 steps for conducting an evaluation](#) and [other evaluation resources](#) for its evaluators. If this is not the case with your organization, I strongly suggest that you rely on a widely used and accepted framework. This will ensure that those sponsoring the evaluation have confidence in your process and results. In turn, they will be more likely to use the results from your evaluation.

Evaluation Standards

In addition to following a widely used evaluation framework, such as those provided by the WHO, CIHR, or CDC, it is also important to follow certain ethical standards when conducting your evaluations. Remember, just because a program is loved and supported does not mean that it is meeting its goals. By committing yourself to certain evaluation standards you can ensure that your evaluation will merit the attention of decision makers who are entrusted with the governance of our health systems.

As you can imagine there are many standards against which you can judge your work, but generally you should ensure that your evaluation (1) meets the

information needs of its intended users, (2) is feasible – meaning realistic, prudent, diplomatic and frugal, (3) is conducted legally and ethically with regard to the welfare of those involved and those affected by its results, and (4) is accurate in its portrayal of the merit and worth of the program being evaluated. In other words, all evaluations should be useful, practical, ethical, and technically sound.

The *Joint Committee on Standards for Educational Evaluations* has identified 30 standards related to these four categories of standards (See **Table 2**). These are provided under four headings of Utility, Feasibility, Propriety, and Accuracy. While not originally intended for public health evaluations, the evaluation standards are useful within healthcare settings. So, regardless of whether these are a formal part of your selected evaluation framework. I would highly recommend that each of these are considered, at least to some degree, to ensure a successful implementation of your evaluation.

Table 2. Evaluation Standards

Utility Standards

The utility standards are intended to ensure that an evaluation will serve the information needs of intended users. These standards are as follows:

- **Stakeholder Identification.** Persons involved in or affected by the evaluation should be identified, so that their needs can be addressed.
- **Evaluator Credibility.** The persons conducting the evaluation should be both trustworthy and competent to perform the evaluation, so that the evaluation findings achieve maximum credibility and acceptance.
- **Information Scope and Selection.** Information collected should be broadly selected to address pertinent questions about the program and be responsive to the needs and interests of clients and other specified stakeholders.
- **Values Identification.** The perspectives, procedures, and rationale used to interpret the findings should be carefully described, so that the bases for value judgments are clear.
- **Report Clarity.** Evaluation reports should clearly describe the program being evaluated, including its context, and the purposes, procedures, and findings of the evaluation, so that essential information is provided and easily understood.
- **Report Timeliness and Dissemination.** Significant interim findings and evaluation reports should be disseminated to intended users, so that they can be used in a timely fashion.
- **Evaluation Impact.** Evaluations should be planned, conducted, and reported in ways that encourage follow-through by stakeholders, so that the likelihood that the evaluation will be used is increased.

Feasibility Standards

The feasibility standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal. The standards are as follows:

- **Practical Procedures.** The evaluation procedures should be practical, to keep disruption to a minimum while needed information is obtained.
- **Political Viability.** The evaluation should be planned and conducted with anticipation of the different positions of various interest groups, so that their cooperation may be obtained, and so that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted.
- **Cost Effectiveness.** The evaluation should be efficient and produce information of sufficient value, so that the resources expended can be justified.

Propriety Standards

The propriety standards are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results. These standards are as follows:

- **Service Orientation.** Evaluations should be designed to assist organizations to address and effectively serve the needs of the full range of targeted participants.
- **Formal Agreements.** Obligations of the formal parties to an evaluation (what is to be done, how, by whom, when) should be agreed to in writing, so that these parties are obligated to adhere to all conditions of the agreement or formally to renegotiate it.
- **Rights of Human Subjects.** Evaluations should be designed and conducted to respect and protect the rights and welfare of human subjects.
- **Human Interactions Evaluators.** Should respect human dignity and worth in their interactions with other persons associated with an evaluation, so that participants are not threatened or harmed.
- **Complete and Fair Assessment.** The evaluation should be complete and fair in its examination and recording of strengths and weaknesses of the program being evaluated, so that strengths can be built upon and problem areas addressed.
- **Disclosure of Findings.** The formal parties to an evaluation should ensure that the full set of evaluation findings along with pertinent limitations are made accessible to the persons affected by the evaluation, and any others with expressed legal rights to receive the results.
- **Conflict of Interest.** Conflict of interest should be dealt with openly and honestly, so that it does not compromise the evaluation processes and results.
- **Fiscal Responsibility.** The evaluator's allocation and expenditure of resources should reflect sound accountability procedures and otherwise be prudent and ethically responsible, so that expenditures are accounted for and appropriate.

Accuracy Standards

The accuracy standards are intended to ensure that an evaluation will reveal and convey technically adequate information about the features that determine worth or merit of the program being evaluated. The standards are as follows:

- **Program Documentation.** The program being evaluated should be described and documented clearly and accurately, so that the program is clearly identified.
- **Context Analysis.** The context in which the program exists should be examined in enough detail, so that its likely influences on the program can be identified.
- **Described Purposes and Procedures.** The purposes and procedures of the evaluation should be monitored and described in enough detail, so that they can be identified and assessed.
- **Defensible Information Sources.** The sources of information used in a program evaluation should be described in enough detail, so that the adequacy of the information can be assessed.

In considering these standards, you might ask questions about your evaluation plan such as those provided by the CDC in **Table 3**. Other guidelines are also available in the literature. For example, those provided by the *Canadian Evaluation Society* emphasize (a) competence, (b) integrity, and (c) accountability as core standards for evaluators. Meanwhile, the *American Evaluation Society* focuses on (a) systematic inquiry, (b) competence, (c) integrity/honesty, (d) respect for people, and (e) responsibilities for general and public welfare. Undoubtedly, all of these standards are important for evaluators as they navigate the complex environments in which they work.

The standards, I admit, probably seem a bit abstract and overly intuitive to us. However, the challenge of conducting a high quality evaluation is not an easy one. You will undoubtedly encounter dilemmas that test your intuition and ability to do what is best for your evaluation project. For example, imagine you have been invited by a community-based organization to evaluate their youth program aimed at improving mental health and resilience. Now, what would you do if the results came out against the program – either showing a null result or a decrease in resilience and mental wellbeing?

Your credibility as an evaluator will be essential to the successful uptake of your final recommendations, as will the community's trust in you and the process by which the results of your evaluation were obtained. You cannot learn to navigate

this sort of situation from a textbook – but these standards can help you be more conscientious about the work you do.

Table 3. Questions to help you meet evaluation standards

Principal	Questions
Utility	<ul style="list-style-type: none"> • Have key stakeholders who can assist with access to respondents been consulted? • Are methods and sources appropriate to the intended purpose and use of the data? • Have key stakeholders been consulted to ensure there are no preferences for or obstacles to selected methods or sources? • Are there specific methods or sources that will enhance the credibility of the data with key users and stakeholders?
Feasibility	<ul style="list-style-type: none"> • Can the data methods and sources be implemented within the time and budget for the project? • Does the evaluation team have the expertise to implement the chosen methods? • Are the methods and sources consistent with the culture and characteristics of the respondents, such as language and literacy level? • Are logistics and protocols realistic given the time and resources that can be devoted to data collection?
Propriety	<ul style="list-style-type: none"> • Will data collection be unduly disruptive? • Are there issues of safety of respondents or confidentiality that must be addressed? • Are the methods and sources appropriate to the culture and characteristics of the respondents—will they understand what they are being asked?
Accuracy	<ul style="list-style-type: none"> • Are appropriate QA procedures in place to ensure quality of data collection? • Are enough data being collected — i.e., to support chosen confidence levels or statistical power? • Are methods and sources consistent with the nature of the problem, the sensitivity of the issue, and the knowledge level of the respondents?

Source: The Program Evaluation Standards: How to Assess Evaluations of Educational Programs by The Joint Committee on Standards for Educational Evaluation

Learning Activities

1. Skim through the evaluation workbooks and guides provided in the additional readings for this section. Based on the steps outlined in each make your own list of evaluation steps classified into (1) planning-related steps, (2) implementation-related steps, and (3) utilization-related steps. You may have as few or as many steps as you think are needed.
2. Consider the four categories of evaluation standards outlined in Table 2. How are these broad categories of standards related to each other?
3. Again thinking about the four categories of evaluation standards outlined in Table 2, have you ever seen an evaluation that violated one of these standards? Or do you know of a time when maintaining these standards was an important part of the evaluation process?
4. Review the Program Evaluation Standards in Table 2 and compare and contrast them with the Program Evaluation Standards of the Canadian Evaluation Society – which are, in fact, adapted from the standards originally developed by the Joint Committee on Standards for Educational Evaluation.
5. How does Indigenous reconciliation relate to the steps and standards reviewed in this section? Are there additional steps or standards that you think should be considered when working towards reconciliation?

Additional Readings

- World Health Organization. "Evaluation Practice Handbook." WHO.
- Public Health Ontario. "Evaluating Health Promotion Programs: Introductory Workbook."
- Centers for Disease Control and Prevention. "Framework for Program Evaluation in Public Health." MMWR.
- Canadian Institute of Health Research. "A Guide to Evaluation in Health Research."

PRELIMINARY EVALUATIONS

Learning Objectives

- Identify the underlying goals and objectives of programs and evaluations.
- Utilize logic models as a program implementation and evaluation tool.
- Engage stakeholders in the evaluation process.

Part 1

What are goals and objectives?



Source: <https://uat.canadaabroad.com/product/60-minute-consultation-person/>

Goals and Objectives

Consider the following quote:

“Evaluation is an orchestrated set of processes and evidence-based products for improved decision-making, including decisions about accountability

I like this definition of evaluation because of its emphasis on decision-making. You will recall from earlier in this text that we distinguished between research and evaluation based on the emphasis that evaluation places on value. It is your job as an evaluator to determine the value of an intervention, but in order to do so you have to understand the goals and objectives of the program. Knowing these goals and objectives, you will be able to develop the goals and objectives of your own evaluation process. The goals and objectives of your evaluation will guide all of the decisions you make with regards to your evaluation process.

Thus, when we speak about goals and objectives, we are speaking about two types: those relating to the program (i.e., program goals, program objectives) and those relating to the evaluation (i.e., evaluation goals, evaluation objectives).

Before we go any further we should define the terms goals and objectives:

- **Goals** describe the broad aims or purposes of an intervention. Sometimes goals come in the form of an organizational mandate. Other times they are set by key decision makers. Goals are typically long term and wide in scope.
- **Objectives** are the strategies or implementation steps used to attain your identified goals. Objectives are usually set by program managers and implementation specialists. Objectives are more specific than goals and usually achievable in the immediate or near future.

For example, the dramatic and rising costs of long term hospitalizations among those requiring acute care recently encouraged the Premier of Ontario to denounce “Hallway Medicine” – the practice of keeping patients in hallways on transport until other beds become available. The desire to eliminate hallway medicine is a goal. An objective related to this goal might be to increase the number of beds at each major hospital in Toronto and Ottawa by 5% before January 2020. This is an objective.



Figure 2. Relationships between Program Goals and Objectives, Evaluation Goals and Objectives, and Evaluation Methodology

It is important that objectives are well constructed, and as highlighted in Figure 1, are connected to the other activities of the evaluation. This is why planning for an evaluation at the start of a program is important. Ideally, the evaluation of the program should be planned at the same time the program is being designed. Of course, this is not always possible. In any case, it is important that objectives follow the SMART objective framework, which we discuss in the next section.

SMART Objectives

SMART objectives are:

- **Specific:** Well defined, clear, and unambiguous.
- **Measurable:** Include specific criteria that measure your progress towards completion of the goal.
- **Achievable:** Attainable and possible to achieve.
- **Realistic:** Within reach, realistic, and relevant.
- **Timely:** Can be completed in a reasonable time.

Table 4 includes some questions that you might ask about the project being implemented. As you construct goals for both programs and for the evaluation activities that will accompany them, it is important that each of these questions are considered by the decision makers and program managers who are responsible for the project. It is often the job of the evaluator to remind leaders of these principals.

Table 4. Questions for creating SMART objectives

Principal	Questions
Specific	<ul style="list-style-type: none"> • Who: Who is involved in this goal? • What: What do I want to accomplish? • Where: Where is this goal to be achieved? • When: When do I want to achieve this goal? • Why: Why do I want to achieve this goal?
Measurable	<ul style="list-style-type: none"> • How many/much? • How do I know if I have reached my goal? • What is my indicator of progress?
Achievable	<ul style="list-style-type: none"> • Do I have the resources and capabilities to achieve the goal? If not, what am I missing? • Have others done it successfully before?
Realistic	<ul style="list-style-type: none"> • Is the goal realistic and within reach? • Is the goal reachable given the time and resources? • Are you able to commit to achieving the goal?
Timely	<ul style="list-style-type: none"> • Does my goal have a deadline? • By when do you want to achieve your goal?

Learning Activities

1. Write a goal that aims to address each of the following three public health concerns:
 - a. People living in rural communities are less likely to test for HIV.
 - b. Wait times for long term care facilities are too long.
 - c. Indigenous women have elevated breast cancer mortality

2. Write three SMART objectives for each of the following goals?
 - a. Improve recovery times for knee replacement patients.
 - b. Increase survival rate among patients eligible for liver transplant.
 - c. Reduce opioid-related overdose deaths.

3. Which aspect of each of the following objectives should be improved to make it a SMART objective?
 - a. Calgary hospital will hire 3 new public health social workers for the emergency department to manage intake of individuals with acute mental health conditions.
 - b. Vancouver health authority will make more needles and syringes available through community pharmacies by 2020.
 - c. Victoria General hospital will increase the number of nurses employed by 35 to reduce patient load by 2020.

4. Why do you think it is important to understand the goals and objectives of a program prior to developing the goals and objectives of your evaluation plan?

Additional Readings

- Bjerke & Renger. (2017). “Being smart about writing SMART objectives.” Evaluation and Program Planning.

Part 2

What are logic models?



Source: <https://www.johnlocke.org/research/direct-primary-care-restoring-the-doctor-patient-relationship/>

Logic Models

A SMART objective is only one piece of a bigger puzzle when it comes to understanding the scope of a public health policy or program. Logic models are the frameworks we use to fit other pieces of the intervention into place so that we can develop an intervention that reflects the nuances of the program or policy being studied.

So what is a logic model? Well, simply, a logic model is visual depiction of the program's goal, the resources available to execute that goal, the activities or interventions that will be undertaken, the expected outputs or products of the program, and the outcomes and impacts that are expected to result from successful implementation. In some instances, the situation surrounding the intervention, the list of assumptions underlying it, and any anticipated external factors are also included to help contextualize the intervention. Links drawn between objectives, inputs, outputs, and outcomes help ensure continuity and logic of a program. If these links cannot be drawn, then the program will not likely succeed. An evaluation that begins with the analysis of a logic

model can be a sure way of identifying potential challenges and can help focus what aspects of the program need to be evaluated.

Case Study: Primary Child Lunch Education Program

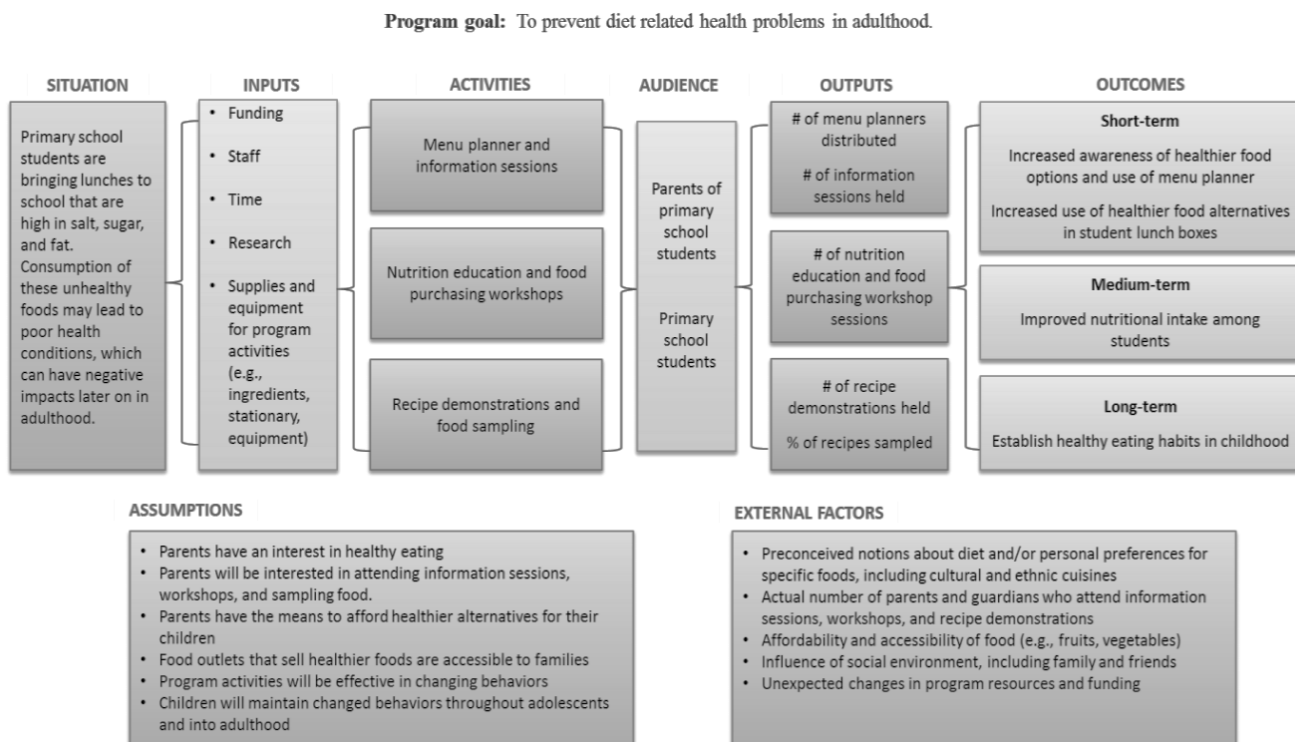
To give you a sense of what a logic model includes, consider a scenario in which you have been asked to create a logic model that describes a school program that aims to improve diet related problems as individuals age. I have provided some of the details for each aspect of a traditional logic model. Please note that not all logic models include all aspects, but each aspect is helpful and I would encourage you to include these as often as possible.

- **Program Goal:** To prevent diet related health problems in adulthood.
- **Situation:** Primary school students are bringing lunches to school that are high in salt, sugar, and fat; Consumption of unhealthy foods may lead to poor health conditions which can have negative impacts later on in adulthood.
- **Assumptions:** Parents have an interest in healthy eating; parents will be interested in attending information sessions, workshops, and sampling food; parents have the means to afford healthier alternatives for their children; food outlets that sell healthier foods are accessible to families; program activities will be effective in changing behaviours; children will maintain changed behaviours throughout adolescents and into adulthood.
- **Inputs:** Funding, Staff, Time, Research, Supplies and equipment for program activities (e.g., ingredients, stationary equipment).
- **Activities:** Menu planner and information sessions; nutrition education and food purchasing workshops; recipe demonstrations and food sampling.
- **Audience:** Parents of primary school students; primary school students.
- **Outputs:** # of menu planners distributed; # of information sessions held; # of nutrition education and food purchasing workshop sessions; # of recipe demonstrations held; % of recipes sampled.
- **Short-term outcomes:** Increased awareness of healthier food options and use of menu planner; Increased use of healthier food alternatives in student lunch box.
- **Medium-term outcomes:** Improved nutritional intake among students.
- **Long-term outcomes:** Established healthy eating habits in childhood.

- **External factors:** Preconceived notions about diet and/or personal preferences for specific foods, including cultural and ethnic cuisines; actual number of parents and guardians who attend information sessions, workshops, and recipe demonstrations; Affordability and accessibility of food (e.g., fruits and vegetables); influence of social environment, including family and friends; unexpected changes in program resources and funding.

The bullet points above do not reflect a logic model. As I said before, logic models are illustrative. In constructing a logic model each of the components I provided would be organized linearly as demonstrated in **Figure 3**.

Figure 3. Example Logic Model format



Key Components of Logic Models

Now that you have an example of what a logic model looks like, let's get some definitions out of the way. Below are definitions for each of the essential components of a logic model:

- **Inputs** are the resources you have at your disposal to direct towards the program. These include the human, financial, organizational, and

community resources that have been allocated for meeting the goal of the project.

- **Activities** are what the program does with the resources that have been allocated. It is the implementation of strategies and projects that will allow you to meet your program objectives. This may include a process, tool, event, a technological innovation, or any other intentional part of the program that is implemented in order to bring about a change or result.
- **Outputs** are the first of three types of results that come from a program and should be considered in a logic model. Outputs are the direct products of a programs activities.
- **Outcomes** are another type of result. These are the specific changes in a target patient’s behavior, knowledge, skills, status, or level of functioning. Many evaluators will break these out into short term (1-3 years), medium term (4-6 years), and long term outcomes (7-10 years). However, these breakdowns are arbitrary and the duration of what is considered “short” or “long” may vary from context to context.
- **Impacts** are the last type of result that comes from a program. Measuring impacts is often difficult since it takes place at a population or sub-population level. Impacts measure the fundamental intended or unintended changes within an organization, community, or health system that result from the implementation of a program. Usually, impacts are strongly and closely aligned with the goals.

The roadmap provided by a logic model can help you understand what essential links need to be evaluated and can help guide your efforts to conduct a thorough evaluation rooted in the intended aims of the program or policy being studied.

Learning Activities

1. Determine whether each of the following are an input, activity, output, or outcome:
 - a. Number of volunteer encounters with patient delivered in the diabetes clinic.
 - b. Provide pharmacist counselling sessions with patients at risk for overdose.
 - c. Number of nutrition workshops organized by the staff health coordinator.

2. Island Health Authority (IHA) wants to meet the needs of residents who are turning to Victoria General Hospital's Emergency Room for mental health care due to lack of mental health coverage. Because care at the hospital is expensive and not the best way to offer care, IHA is working to create a free mental health clinic. Based on that scenario, answer the following questions:
 - a. What is the goal of Island Health Authority in this scenario?
 - b. What resources would you likely need to be dedicated to completing this goal?
 - c. What activities might be undertaken in the process of meeting this goal?
 - d. What outputs should be measured to document whether the activities have been implemented appropriately?
 - e. What outcomes would you want to watch for following the implementation of the program?
 - f. What is the impact that you would expect to see as a result of the implemented program?

3. Based on the questions outlined above, construct a logic model for the Island Health Authority's mental health care clinic.

4. What assumptions do you believe underlie the logic model you created?

5. Review the additional readings and develop in your own words an explanation for why logic models are important in the evaluation process.

Additional Readings

- W.K. Kellogg Foundation. "[Logic Model Development Guide.](#)"
- Peyto & Scicchitano (2017). "[Devil is in the details: Using logic models to investigate program process.](#)"
- Miller (2013). "[Logic models: a useful way to study theories of evaluation practice?](#)"
- Ball et al. (2017) "[Using logic models to enhance the methodological quality of primary health-care interventions.](#)"

Part 3

How to involve stakeholders?

Source: <https://www.michigangastro.com/patient-forms>

Roles and Responsibilities

Even as you begin to plan an evaluation, it should be abundantly clear that you cannot do an evaluation alone. There are myriads of stakeholders who will have to be engaged, at least to varying extents, as you proceed in evaluating any given policy or program. The broader the impacts of an evaluation, the greater number of stakeholders will need to be engaged. Generally speaking, the list below provides an accounting of stakeholders that should be considered when planning and conducting research and evaluation activities:

- **Public:** Individuals not directly interested in the intervention.
- **Patients and Communities:** The individuals who are affected by an intervention, as well as their advocates.
- **Subject Matter Experts:** The staff or other specialists who are experts on the specific program, the relevant topic areas, or the theoretical framework guiding the program.
- **Industry:** Corporate organizations whose interests might be impacted by changes in policy or program.

- **Analysts:** Researchers or policy specialists who advise decision makers and other program leaders.
- **Program Manager:** The persons responsible for implementing the program.
- **Advisory Committee:** A group of stakeholders who might be convened to provide advice or guidance to the evaluation team.
- **Decision makers:** The person or group accountable for making decisions about the program.
- **Sponsor:** The individual or organization responsible for administering and overseeing the evaluation and receiving its findings.
- **Evaluation team:** The team members assigned to conduct the evaluation.

Individual stakeholders who are necessary for the evaluation to move forward will often create a terms of reference document to facilitate the sponsor-evaluator relationship. **Terms of reference** are a formalization of the evaluation plan and serve as a guide for the parties coming together to conduct the evaluation. The initial draft of the terms of reference will often come from the sponsor and will offer an explicit, focused mandate for the evaluation team. These terms will outline what is being evaluated, why, who should be evolved, and the expected outputs. The United Nations Evaluation Group (UNEG) suggests that sponsors and evaluators use the quality checklist in Table 5 to ensure that the terms of reference include all the necessary details for the evaluation take place.

Table 5. UNEG Quality Checklist for Evaluation Terms of Reference Reports

<p>1. Evaluation Purpose</p> <ul style="list-style-type: none"> • 1.0 The Terms of Reference (TOR) specifies the purpose of the evaluation and how it will be used. • 1.1 The TOR references the mandate for the conduct of the evaluation. • 1.2 The purpose of the evaluation identified in the TOR clearly states why the evaluation is being done, including justification for why it is being done at this time. • 1.3 The TOR identifies the primary and secondary audiences for the evaluation and how the evaluation will be useful.
<p>2. Evaluation Objectives</p> <ul style="list-style-type: none"> • 2.0 The TOR includes clearly defined, relevant and feasible objectives. • 2.1 The evaluation objective(s) clearly follow from the overall purpose of the evaluation. • 2.2 The TOR evaluation objectives are realistic and achievable, in light of the information that can be collected in the context of the undertaking.

3. Evaluation Context

- 3.0 The TOR includes sufficient and relevant contextual information.
- 3.1 The TOR adequately describes the particular political, programmatic and governance environment in which the evaluation will be taking place. For example, the most relevant aspects of the economic, social and political context are described.
- 3.2 The TOR adequately describes the most relevant programmatic and/or thematic aspects relevant to the evaluation.

4. Evaluation Scope

- 4.0 The TOR includes the scope of the evaluation.
- 4.1 The TOR explicitly and clearly defines what will and will not be covered, including, for example, the timeframe, phase in the project and/or geographical area to be covered by the evaluation.
- 4.2 The scope of the evaluation is adequate to meet the stated evaluation objective(s).
- 4.3 The scope of the evaluation is feasible given resources and time considerations.

5. Evaluation Criteria

- 5.0 The TOR specifies the criteria that will be utilized to guide the evaluation.
- 5.1 The TOR specifies the evaluation criteria against which the subject to be evaluated will be assessed, including, for example, relevance, efficiency, effectiveness, impact and/or sustainability.
- 5.2 The TOR spells out any additional criteria of relevance to the particular type of evaluation being undertaken, such as evaluations of development, humanitarian response, and normative programmes.
- 5.3 The scope of the evaluation is feasible given resources and time considerations.

6. Tailored Evaluation Questions

- 6.0 The TOR includes a comprehensive and tailored set of evaluation questions within the framework of the evaluation criteria.
- 6.1 The TOR contains a set of evaluation questions that are directly related to both the objectives of the evaluation and the criteria against which the subject will be assessed.
- 6.2 The set of evaluation questions adds further detail to the objectives and contributes to further defining the scope.
- 6.3 The set of evaluation questions is comprehensive enough that they raise the most pertinent evaluation questions, while at the same time being concise enough to provide users with a clear overview of the evaluation's objectives.
- 6.4 Factoring in the information that will be collected and the context of the evaluation, evidence backed answers to the set of evaluation questions is achievable.

7. Methodology

- 7.0 The TOR specifies the methods for data collection and analysis, including information on the overall methodological design
- 7.1 The TOR contains a clear and accessible methodological plan. Preferably, a standalone section that is clearly delineated from other information contained in the TOR.
- 7.2 The TOR states the overall methodological approach and design for the evaluation. Examples of approaches include participatory, utilization-focused, theory-based and gender and human rights responsive. Examples of overall design include non- experimental, quasiexperimental and experimental.
- 7.3 The data collection and analysis methods in the TOR are sufficiently rigorous to assess the subject of the evaluation and ensure a complete, fair and unbiased assessment. For example, there will be sufficient data to address all evaluation questions.
- 7.4 The evaluation methodology includes multiple methods (triangulation); preferably with analysis of both quantitative and qualitative data and with a range of stakeholders covered by the data collection methods. 7.5 Logical and explicit linkages are provided between data sources, data collection methods and analysis methods. For example, sampling plans are included.
- 7.6 The evaluation methodology takes into account the overall purpose of the evaluation, as well as the needs of the users and other stakeholders.
- 7.7 The evaluation methodology explicitly and clearly states the limitations of the chosen evaluation methods.
- 7.8 The TOR specifies that the evaluation will follow UNEG norms and standards for evaluations, as well as ethical guidelines.

8. Evaluation Work Plan

- 8.0 The TOR includes a work plan
- 8.1 The TOR work plan states the outputs that will be delivered by the evaluation team, including information on the degree to which the evaluation report will be accessible to stakeholders, including the public.
- 8.2 The TOR work plan describes the key stages of the evaluation process and the project time line.
- 8.3 The TOR work plan establishes clear roles and responsibilities for evaluation team members, the commissioning organization and other stakeholders in the evaluation process.
- 8.4 The TOR work plan describes the evaluation quality assurance process.
- 8.5 The TOR work plan describes the process, if any, for obtaining and incorporating evaluand comments on a draft evaluation report. 8.6 The TOR work plan includes an evaluation project budget.

9. Gender and Human Rights

- 9.0 The TOR specifies how a human rights and gender perspective will be incorporated in the evaluation design.
- 9.1 The TOR indicates both duty bearers and rights holders (particularly women and other groups subject to discrimination) as primary users of the evaluation and how they will be involved in the evaluation process.
- 9.2 The TOR spells out the relevant instruments or policies on human rights and gender equality that will guide the evaluation process.
- 9.3 The TOR includes an assessment of relevant human rights and gender equality aspects through the selection of the evaluation criteria and questions.
- 9.4 The TOR specifies an evaluation approach and data collection and analysis methods that are human rights based and gender sensitive and for evaluation data to be disaggregated by sex, ethnicity, age, disability, etc.
- 9.5 The TOR defines the level of expertise needed among the evaluation team on gender equality and human rights and their responsibilities in this regard and calls for a gender balanced and culturally diverse team that makes use of national/regional evaluation expertise.

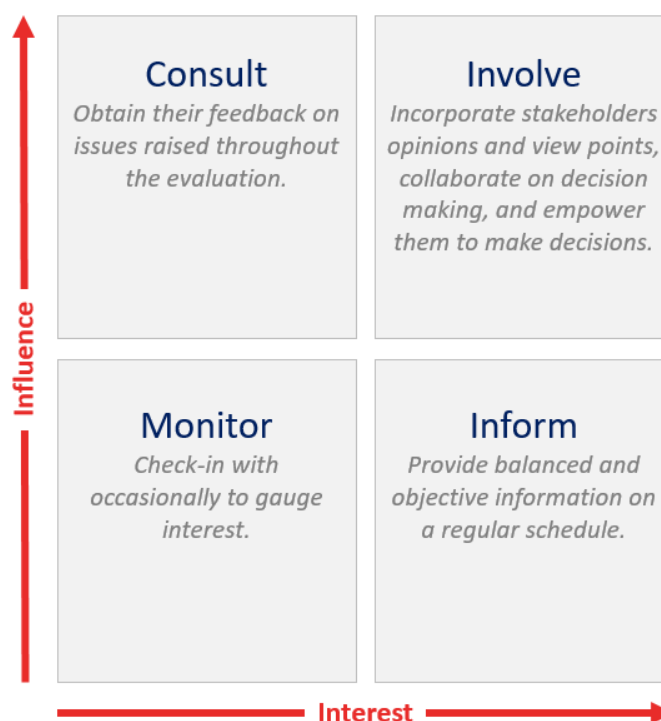
As you can tell, the terms of reference can be quite specific and developing them will require collaboration and cooperation between yourself as an evaluator and the sponsors of the evaluation. Involving stakeholders at this stage will ensure that the evaluation that proceeds meets the demands of all those involved.

Stakeholder Analysis

When engaging stakeholders, stakeholder analysis is one of the helpful tools you can turn to in order to determine how stakeholders should be engaged. After all, if all the stakeholders of a project were involved in all decisions – the process of evaluation would move along exceedingly slow. While there are many stakeholder engagement strategies, one form of analysis involves assessing how *interested* each stakeholder is in the program and its evaluation and then determining how *influential* they are. As you can see in Figure 4, the intersection of influence and interest can guide how you will engage with stakeholders. Individuals or groups that are highly influential, but not very interested should be consulted. Individuals or groups that are not influential and not interested should be monitored for interest. Individuals who are interested but not influential should be informed, and individuals who are interested and influential should be involved, collaborated with, and empowered. It is your job as an evaluator to ensure that the stakeholders are identified and classified appropriately – however, you will

likely rely on the advice of many stakeholders in creating a community of practice that is willing to work together to complete your evaluation. It is often helpful to keep track of stakeholders, their representatives, their assigned level of involvement, their contact details, and any important notes about the stakeholders and their interests in a single document. This record is sometimes referred to as a stakeholder table. A regular review of the stakeholder table, as well as procedures around when various stakeholders will be contacted will help ensure that stakeholders are appropriately engaged in the evaluation process.

Figure 4. Stakeholder Analysis Grid



Participatory Evaluation

In recent years, evaluations have become increasingly participatory – mirroring the emergence of community-based participatory research (CBPR) in the academic world. Patton (2008) has identified nine principles of participatory evaluation that can be used when conducting an evaluation that values the participation of stakeholders. These principles are as follows:

- The process involves participants in learning skills.
- Participants own the evaluation and are active in the process.

- Participants focus the evaluation on what they consider important.
- Participants work together as a group.
- The whole evaluation process is understandable and meaningful to the participants.
- Accountability to oneself and to others is valued and supported.
- The perspectives and expertise of all persons are recognized and valued.
- The evaluator facilitates the process and is a collaborator and a resource for the team.
- The status of the evaluator relative to the team is minimized (to allow for equitable participation).

Participatory evaluations are now becoming the standard for evaluation methodology. It is therefore important that evaluators be prepared to work with individuals from an array of beliefs, traditions, and backgrounds. Demonstrating cultural competence and an enthusiastic desire to work with stakeholders will help you gain their trust and support. According to the United Nations Population Fund, you can do so by:

- Investing time in knowing the culture in which you work,
- Listening to what the community has to say,
- Demonstrating respect,
- Being inclusive,
- Honoring commitments,
- Finding common ground,
- Building capacity in stakeholders,
- Support individuals in doing what they do best,
- Providing solid evidence for them to act on, and
- Promoting objectivity over your personal judgement.

The ways you engage with stakeholders will likely vary greatly depending on whether you are an internal or external evaluator. Generally, speaking internal evaluations provide a better framework for evaluation. However, it is not always possible to have an in-house evaluation team. Under some circumstances you might even want the benefits of bringing in an external evaluator. If you happen to be an external evaluator, building bridges with stakeholders is all the more important.

Evaluation Ethics

One additional note to consider when working with or as an external evaluator is the role of research ethics boards in evaluations. Research ethics boards (REBs) vary in how they perceive their role in evaluation. Therefore, it is important to contact your REB prior to engaging in an evaluation. In some cases, when external funding is received, an ethics review may be necessary. Additionally, when evaluation findings will be used for more than quality improvement of programs, it may be necessary to secure REB approval prior to publication of results. In any case, regardless of whether REB approval is needed, the following considerations drafted by the Canadian Institutes of Health Research, should be made:

- **Negative impacts on staff:** There is a risk that difficulties identified with a program may be attributed to specific staff. Negative findings may also impact staff, including in situations where a program is shut down following an evaluation showing inefficacy.
- **Negative impacts on patients/clients:** Clinical and system redesign interventions can have a negative impact on patients and families.
- **Negative impacts on organization:** Negative evaluations can have a significant impact on an organization, including loss of program funding. It is important to recognize that the role of an evaluator is to report honestly and openly with the evaluation sponsor.
- **Opportunity costs:** Inadequate or limited evaluation can result in continuing to provide resources to a lackluster service, meaning that other initiatives cannot be funded.

As we discussed earlier, standards that support high quality evaluations will help ensure that the evaluation team can act ethically. Furthermore, terms of reference can be used to guide program evaluators to ensure that negative impacts of evaluation are mitigated or accounted for prior to the initiation of the evaluation. The duty of an evaluator to their professional ethics, to the stakeholders involved, and to the quality of the evaluations they conduct should be at the forefront of an evaluators work as they seek to improve the quality of community-level programs and health systems policies.

Learning Activities

6. Check out the following list of stakeholders and determine which of them should be monitored, consulted, involved, or informed on a sex education program administered by Vancouver Coastal Health? After classifying them, consider why you made the decision you did:
 - a. [Vancouver Coastal Health](#)
 - b. [Youth Co](#)
 - c. [Sex Education is our Right](#)
 - d. [Healthy Schools BC](#)
 - e. [LifeLabs](#)
 - f. [Community Based Research Centre](#)
 - g. [Vancouver Division of Family Practice](#)
 - h. [North Shore Division of Family Practice](#)
 - i. [Ministry of Health](#)
7. Imagine you are the director of the emergency room at [St. Paul's Hospital](#) in Vancouver and you are preparing for the [hospital's relocation](#). You want to ensure that all the stakeholders in the hospital's emergency services are prepared for the transition. Which stakeholders would you contact and why?
8. Assume you are conducting a quality improvement evaluation at Island Health. According to their [policies](#), would you need formal review by a research ethics board to proceed with your study?

Additional Readings

- Goodman & Thompson (2017). "[The science of stakeholder engagement in research: classification, implementation, and evaluation.](#)"
- Fiscella et al. (2015). "[Ethical oversight in quality improvement and quality improvement research: new approaches to promote a learning health care system.](#)"
- Morris & Cohn (1993). "[Program Evaluators and Ethical Challenges: A National Survey.](#)"



Part 1

What are Preliminary evaluations?

Source: https://www.thestar.com/news/gta/2011/10/02/a_creative_refuge_in_the_city.html

Preliminary Evaluations

If you recall from earlier in this text, we described preliminary evaluations as those which take place prior to the development of a new intervention. In reality, there may or may not already be a process, program, or policy in place addressing a specific goal. With a preliminary evaluation, however, you are aiming to understand whether the needs of individuals, organizations, and communities are being met irrespective of whether or not there are already resources dedicated to addressing those needs.

The central topics addressed by preliminary evaluations are captured by questions such as the following:

- “What is being done?”
- “What should we be doing?”
- “What factors will impact what we will do?”

However, I want to stress that simplistic and off-the-cuff answers to these questions are rarely sufficient. Intuition and even expertise are no match for a careful and systematic investigation of an issue. Indeed, contemporary health care calls us to engage with diverse stakeholders who often have conflicting interests and values, across a myriad of complex and nuanced health conditions with limited resources and distracted

collaborators. Cooperation across all levels of the health system, from patients and providers to policy makers and politicians, is necessary to implement even the most evidence-supported interventions you can imagine. The challenge is all the more complicated, when what should be done, or even what is being done, is unknown.



Figure 5. General Framework for Program Planning, Implementation, and Evaluation

This is why all policy and program work within the healthcare system should begin with a thorough *assessment* (See **Figure 5**).

Indeed, consider for a moment the skyrocketing number of overdose-related deaths across Canada in the past four or five years. The challenge with these deaths is that there is disagreement over what the fundamental problem even is – which has contributed to the lackluster government response to addressing these preventable deaths. Is the problem the availability of addictive drugs? Or merely poorly manufactured drugs with highly variable titrations of fentanyl? Is it lack of naloxone? Or lack of training around emergency overdose procedures, such as how oxygen is provided? Is it restricted access to safe injection sites or opioid agonist therapies? Or is it the stigma that leaves potential clients uncomfortable with accessing traditional services? In reality, all of these factors contribute to the overdose epidemic and there are probably also many other issues that I have not even come close to touching on.

The challenge for public health leaders is to not only diagnose which problems contribute the most, but to know what can feasibly be done to address the identified problems. In other words, public health leaders must make decisions. The purpose of the preliminary evaluations is thus to help us identify which decisions need to be made, understand what factors limit our decision making capabilities, and consider the full scope and breadth of potential programs and policies.

I wish I could say that there was some universal grand strategy for how these sorts of evaluations could be conducted. Unfortunately, there is not. Preliminary evaluations will look different in different contexts and in addressing different health conditions. Some

organizations will prefer one style of preliminary evaluation over another. Whatever formative strategies are employed, I would urge you to always consider the principles of evaluation reviewed in earlier sections of this text.

Now, I want to provide you with a few tools that may or may not be employed in the course of a formative preliminary evaluation. In general, I have chosen the most widely used methodologies and will outline the general steps that I think best characterize these methods. Of course, various authors and texts will present slightly varied approaches. If you come across these or search them out, I would not worry too much about the differences in evaluation strategy.

What I do want you to pay attention to is the varied evaluation strategies available to you and which ones you think are most helpful.

Document or Data Review

The first evaluation methods that you might turn to when familiarizing yourself with a new evaluation project is to conduct a document review. In general, this evaluation technique is unobtrusive, relatively inexpensive, and provides a behind the scenes view of an issue. According to the US CDC,

“[A] document review is a way of collecting data by reviewing existing documents. The documents may be internal (such as records of what components of an asthma management program were implemented in schools) or external (such as records of emergency room visits by students served by an asthma management program). Documents may include reports, program logs, performance ratings, funding proposals, meeting minutes, newsletters, and marketing materials.”

The document review process allows you to gather relevant background information about the context of the evaluation, understand the organizations goals, intentions, and plans and how these plans align with what is currently being done, to help you understand what existing sources of information are available to you in any forthcoming evaluation, and to understand basic parameters of a program, such as past service delivery benchmarks or costs dedicated to addressing a concern.

The steps for a document review are provided in Table 6.

Table 6. How to plan and conduct a document/data review

Step 1: Identify available documents

Find out what types of documents exist and determine which are most relevant.

Step 2: Gain access to documents

Some documents may require the permission of others before being reviewed. You may need to work with legal experts or other data custodians in your agency to understand what limitations you may face in accessing these documents.

Step 3: Anonymize documents

If you need to review documents that involve confidential data about individuals, develop a system that ensures confidentiality of individual-level data. Developing these processes and guidelines may also help you in securing access to sensitive or confidential documents.

Step 4: Compile relevant documents

Once you have secured access to the documents you need to answer your evaluation questions, compile the documents. It is important that you limit your review to only those documents that answer your evaluation questions.

Step 5: Understand the documents

You will need to talk to the people who know something about the documents you are compiling to better understand the context for which they were developed. This is critical to gathering usable information for your evaluation.

Step 6: Establish accuracy of information

Determining the accuracy of the documents may involve comparing the documents that contain similar information, checking the documents against other data you have collected, and speaking with people who were involved in the development of the documents.

Step 7: Summarize relevant information

Create a data collection form to summarize data gleaned from your document reviews. You may want to include on the form the type of document you are reviewing; a way to reference each document; and information that answers each applicable evaluation question.

Step 8: Analyze your data/document summary

After you have your data summary prepared, there are many options for evaluation. Which you choose will largely depend on what information you collected in your summary form and what you were hoping to learn.

Source: CDC (2018). Data Collection Methods for Evaluation: Document Review

While a document review has plenty of advantages, including those mentioned above, it is important to recognize that existing documentation may be biased through

duration. Furthermore, the undertaking of pulling together and sorting through relevant data sources can be laborious at times – and even with careful curation, the information included in documents may be incomplete or inaccurate.

Literature Reviews

In addition to reviewing the documents of an organization, it is also worth your while to review what is already known about the problem at hand and the potential interventions that can be used to address it. Knowing your organization well will help you to identify from the literature what interventions and findings from the literature are helpful and appropriate for your situation and context. Having evidence to support your recommendations will add credibility to your proposals and are a great way to begin a working relationship. However, it should be remembered that your stakeholders are likely better informed than the literature often is. That said, looking for literature reviews on an issues, or conducting a literature review yourself, can allow you and your stakeholders to engage with the empirical evidence in order to help you better understand what may work for your given situation.

Two great databases for evidence summaries are [The Canadian Agency for Drugs and Technologies in Health \(CADTH\)](#) and the [Cochrane Library](#). You can also search [PubMed](#) for reviews – though reviews from other data sources can be of varying quality. You should also become familiar with [the Institute for Health Metrics and Evaluations \(IHME\)](#), [the Canadian Institute for Health Information \(CIHI\)](#), and [the Centre for Health valuation and Outcome Sciences](#) – each of these provide information and support that will be relevant to how you plan your work. Many topic areas also have research organizations dedicated specifically to their work. For example, in the substance use field, there are a few big key research groups that you can turn to for help and advice: [the Canadian Centre on Substance Use and Addiction \(CCSA\)](#), [the Centre for Addictions and Mental Health \(CAMH\)](#), [the British Columbia Centre on Substance Use \(BCCSU\)](#), and [the Canadian Institute for Substance Use Research \(CISUR\)](#).

If you end up conducting a review for yourself, you have a variety of reviews to choose from – narrative reviews, scoping reviews, systematic reviews, and so on. These generally vary in how rigorous they are conducted – with systematic reviews and meta-analysis being the most technically demanding. The type of review you need to conduct will largely depend on what literature is already out there, what type of research

question you are asking, and how the indicators relevant to your work are traditionally measured.

As you decide on the type of review you will conduct, it is important that you begin to develop your research question. It is important that your research topic is sufficiently narrow as to limit the number of articles included in your review, but not so limited that you don't have enough studies to synthesize. PICOS is one of the widely used frameworks for developing research questions for quantitative studies (Aslam & Emmanuel, 2010) and SPIDER is a widely used tool for developing qualitative research questions (Cooke, Smith, & Booth, 2012).

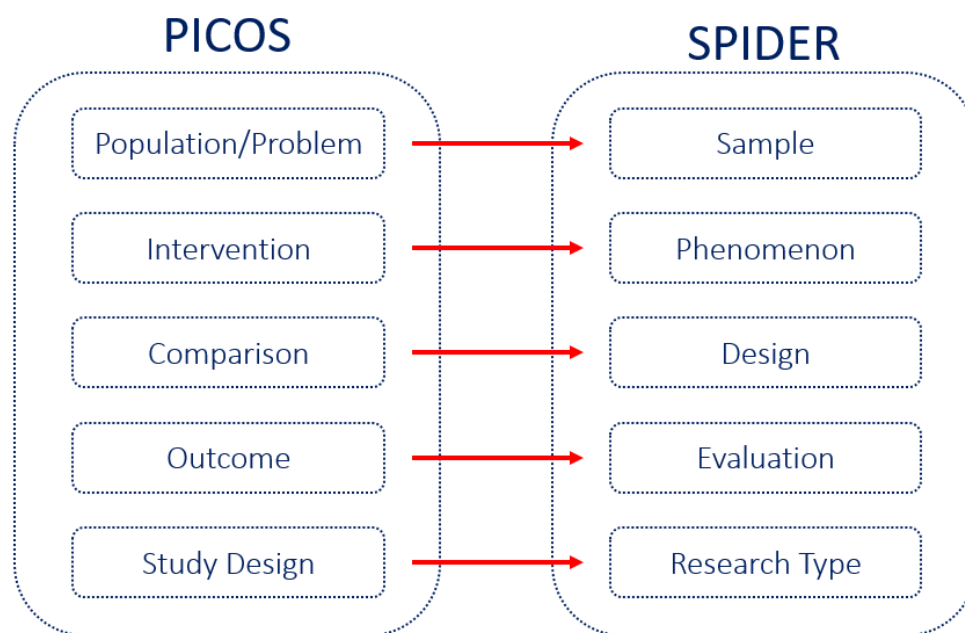


Figure 6. PICOS and SPIDER Elements of Research Questions

Based on the PICOS or SPIDER framework, you will need to identify (1) identify the keywords and keyword combinations you will use to identify potentially relevant studies and (2) establish clear inclusion and exclusion criteria that you will use to screen the results from your search. These search criteria will be used in connection with a database such as *PubMed* or *Web of Science*. Using Boolean operators, you will need to execute your search strategy. Please note that the Boolean operators work slightly different in each platform and you may have to do some googling/searching to figure out how they work. After you have conducted your keyword search, you should export

the list of retrieved articles from PubMed and Web of Science into excel. PubMed provides a tutorial on how you can export the results of your search [here](#).

After you have all the studies in a single excel file, you should first review the study titles. You can create a column that indicates whether or not the title is relevant and mark “yes” or “no” for each study as you review them. At this point in your search, you should err on the side of inclusion if the title does not clearly rule in or out specific studies. In excel, you can then sort your searches by the newly created column and begin to review each abstract to see if you can identify whether the studies not already excluded meet any of your exclusion criteria. I recommend using a series of columns with each column identifying one of your exclusion criteria. In each column you can mark whether the study should be excluded by putting a “1” in the column associated with the criteria it failed to meet. For any studies that you are not able to exclude based on a reading of the manuscript, you may have to read the full manuscript to determine whether it meets your inclusion criteria.

After you have a list of articles, you may decide that you want to further restrict your articles with an additional inclusion or exclusion criteria (e.g., language, location, year of study, study design type, method of outcome measurement). Once you have figured out which articles are included, it is recommended that you create a [PRISMA flow diagram](#) detailing the identification, screening, eligibility, and inclusion of studies into your review (Moher et al., 2015).

The next step is data extraction. You will need to extract the relevant information from each section. Data extraction tools can be set up as individual sheets or as excel tables, with each piece of extracted information as its own column. For example, you will probably want to extract the authors, dates of recruitments, locations, sampling methods, outcome measures, and explanatory measures for each study, as well as general summation of what the study found.

After you have extracted the relevant data, you can critically appraise each study using a tool such as the Joanna Briggs Institute’s [Critical Appraisal tool](#) (Joanna Briggs Institute, 2017). You should also use something like the [PRISMA checklist](#) to write your literature review (Moher et al., 2015) to make sure you are including all the necessary elements of your study. Further, if you are not already using a reference manager, such as [Zotero](#) to manage your references, I highly recommend that you do so. Finally, don’t be

intimidated to look up resources for systematic reviews to help remind you of these steps or to find ways to make these easier:

- <http://guides.library.cornell.edu/c.php?g=459012&p=3137889>
- <http://sites.cardiff.ac.uk/curmed/files/2014/10/NSAMR-Systematic-Review.pdf>
- https://upoj.org/wp-content/uploads/v23/C_GENERAL_Yannascoli.pdf
- <https://www.ccace.ed.ac.uk/research/software-resources/systematic-reviews-and-meta-analyses>

Jurisdictional Scans

Another type of review often conducted in the lead up to an implementation project or evaluation is a cross-jurisdictional scan. These are particularly popular for population-level interventions and policies, but can be conducted at any level. Similar with literature and document reviews, the focus of a jurisdictional scan is on what is being done or has been done in other places. The main goal is usually to identify best practices and consider how problems and solutions are framed or addressed in other settings. Evidence of success in another area, while not a guarantee, is good evidence for potential success in your setting. Shopping around to see what methods and interventions have been used will help decision makers understand better whether the course of action they are considering meets the standards in the field.

One of the key questions you will ask when conducting a cross-jurisdictional scan is what comparison jurisdictions would you like to include. For example, will you include data from each province, from other countries, from only cities with similar demographics, economies, and budgets as your own? Comparing apples to apples can help you know what is realistic or feasible. However, sometimes apples-to-oranges comparisons are useful for framing an issue and promoting change. Many provincial policy scans make comparisons to every province. For example, the availability of prescription drugs through drug formularies is often reviewed by looking what other jurisdictions have made available. You will find that provinces that lead the way in making change tend to have a small cohort of provinces ready to follow them in lock step. One reason for this is that program designers and policy makers tend to be a bit reticent to try something new – particularly if they are in a position where they are held accountable for the success or failure of a program.

The main difference between a jurisdictional scan and other literature or document review procedures is that it explicitly presents data across jurisdictions and it usually focuses on a mixture of academic and gray literature. Because jurisdictional scans can be used in such a wide variety of situations, there is no widely agreed upon methodology for cross-jurisdictional scans. That said, it is generally agreed that the data collection methods should be standardized across each jurisdiction included in your review.

Observation

Similar to cross-jurisdictional scans and document, data, and literature reviews, another relatively inexpensive evaluation method that can help you understand a health program or system better is *observation*. Observation can include a range of activities from sitting in on a committee meeting and watching how committee members interact to shadowing a nurse on their rounds to understand what their work entails.

When planning an observation there are a variety of things for you to consider, which I outline below:

- **Ethics** of observation should always be considered, both for the safety of the observer and those being observed. Suppose, for instance, you happen to observe conflict or unethical behavior. You will need to decide how you will handle these situations and how your decisions might impact your ability to conduct future evaluations. Indeed, the ethics of observation are not only limited to watch or not to watch, but include legitimate questions about how you will act (or not act) as an observer.
- **Emotional stress** can also result from observation, particularly when observing settings in which others are under distress – such as hospital emergency rooms or clinics serving vulnerable populations.
- **Timing** is a major concern in an observation. How frequently does the thing being evaluated occur? How long will it take to observe the entire process? How much observation time is required? How many times do you need to see something done? All of these questions are important to consider when undertaking an observational study.
- **Fatigue** can be a serious burden in observation studies. It can be tedious and tiring.

- The **Hawthorne effect** is a source of confounding bias that causes those being observed to act differently while being observed compared to when they would not be observed. You will need to consider how obtrusive you will be and how you will approach the observational experiences comfortably. If participants have not been already told they are being observed, you will need to determine what you will tell them you are doing, if they ask.

With these challenges at the forefront of your mind, it is useful to recognize that observation represents a spectrum of activities. Krueger ([2017](#)) notes the following 5 degrees or levels of observation:

- Evaluator observes without being noticed.
- Evaluator uses unobtrusive observation (observe but not in an obvious manner) and doesn't interact with participants.
- Evaluator has limited interaction, intervening only when further clarification of actions is needed.
- Evaluator uses active control over the observation, as in the case of a formal interview, to elicit specific types of information.
- Evaluator fully participates in the situation with either a hidden or known identity

Which of these options you choose will largely depend on what you are trying to observe and your relationships with those being observed. Each will help you understand the context of what is being evaluated to a different extent.

Once you have decided what your role will be as an evaluator, you will need to begin to think through what your observation will aim to help you understand. The steps provided in **Table 7** will help you think through a systematic approach for conducting your observation. It is important to know that the steps listed may or may not be conducted in the order listed.

Furthermore, it is useful to remember that a systematic approach does not necessarily need to be a static one. You will find that flexibility and dynamism is needed in conducting most evaluation activities. As you conduct your observations, you will find that you undoubtedly had not considered observing certain important factors that you would like to add to your observation. Do not worry if this is the case.

Table 7. How to conduct an observation.

Step 1: Plan your observation

When planning an evaluation, be sure to look in the literature and see how similar evaluations have been conducted. Don't be afraid to copy and adapt what has been used before for your own project. As you review the literature, you will be establishing a framework or idea about what you are observing. In doing this it is extremely useful to consult with others. Put your thoughts on paper and share with others. Incorporate their feedback.

Step 2: Pilot and refine your observation strategy

Once you have a draft of your observation plan, it is helpful to test pilot your strategy. For instance, if you are looking at the development of a nurse-led intervention in an clinical setting, you might have wanted to count how frequently a nurse reported to their supervising physician or asked them a question. Upon observation, you might find out that it is also important to keep track of the number of times a doctor intervenes on their own initiative. Likewise, don't forget to make notes about what is *not* happening – sometimes things that don't occur are just as important to think about as those things which do. Hopefully, the pilot observation will help you identify additional indicators that should be recorded throughout your evaluation.

Step 3. Finalize your observation checklist and data collection tools

Observational evaluations are not aimless. Rather, they must be planned and thought through. An evaluation checklist is a tool to help focus your evaluation. While you must be careful to ensure that this tool does not prevent you from seeing other important, the checklist helps ensure that the observation is rigorous and consistent. A checklist might include, the duration of activities, frequency of occurrence, the extensiveness and intensity or key activities, demographics of the participants, environmental factors, sequences off events, and behaviors. You might also leave space for visual notes: maps, charts, flow diagrams, time logs, etc.

Step 4: Check for consistency

If you are working with multiple evaluators, you will want to make sure that there is strong agreement between reviewers. Likewise, if you are observing multiple subjects or interactions, comparing across these can help you understand how consistent specific observations are. Ensuring that your observations are appropriately sampled and are representative of what is actually done is an important step to making sure your observational data are useful.

Step 5: Analyze your results

The first step in analyzing your data is to summarize your results. In doing so, you should review your field notes and look for key patterns, themes, and concepts that seem relevant to what you are trying to observe. You might try to create concept maps or diagrams, or a short executive summary, to help you articulate everything you observed or.

Community Mapping

One challenge with observation and document review is they tend to be detached. There is a growing movement in evaluation to incorporate participatory methods in the evaluation process. Participation from individuals on the front line, with expertise that comes from lived experience, is an invaluable resource to an evaluation.

One participatory method that has gained increasing popularity is that of community mapping. Hundreds of publications and reports using this methodology are now in circulation. Generally speaking, community mapping is a diverse methodology, but generally involves four steps (See Table 8). According to Corbett ([2009](#)) most participatory mapping projects aim to:

- communicate spatial knowledge to outsiders,
- aggregate and archive local knowledge,
- inform resource management decisions,
- identify deficits and advocate for investment, and
- increase capacity within communities.

Each mapping project will be unique, but these six general categories provide a focus and direction that will help you decide what data should be collected, how it should be collected, and to whom the results of your mapping project should be shared (e.g., is it for internal use or can the list of locations be compiled and provided as a central resource for the community).

The most novel aspect of the community mapping process, is the participatory way in which the social geography of a region is identified. Maps can be created online using software's designed for marking, by placing stickers on a map, or with markers (See Figure 6). Participants might be the ones drawing the maps, or they might be directing an artist on what to draw. One of the great advantages of community mapping protocols is the ability to adapt the protocol to meet your needs and the level of engagement or readiness among your participants.

Irrespective of what data you are collecting, community mapping activities provide you with an opportunity to connect with individuals – either in focus group settings, one-on-one, or via online surveys. Collecting information from individuals with lived experience will ensure that the program and evaluation activities you implement later on down the

road are informed by those who are most directly impacted by your work. When possible, participatory evaluation activities should be used to facilitate relationships between you and others involved. As such, it is important that all individuals who participate in your assessments feel valued as equal partners on the project. Establishing clear roles and goals can help organize your relationships and give people the expectations they need to succeed as collaborators on your project. This is particularly vital when participants in come from diverse backgrounds or are accustomed to operating within strict hierarchies.



Figure 6. Examples of Community Maps

Table 8. How to conduct a community mapping study

Step 1: Identify goals and Participants

Pre-mapping involves setting the goals, preparing questions, and putting together a team of participants in the map making process. Goals will help you develop the questions that will guide participants as they participate in the mapping project. Finally, it is vital that the right people are around the table when constructing a map. Consider whose perspectives you are getting and whether other people at the table would offer different perspectives. Also consider how individuals engaged in the mapping process might be impacted by the presence of other team members.

Step 2: Identify what will be mapped

Goals will help you identify what aspects of the community you are interested in mapping. For example, you might be interested in where police patrol, or where pharmacies are located, or how people get from one place to another. You will need to reach consensus on what geographic measures are important to your specific research question.

Step 3: Collect Data

The process of data collection will change based on who is at the table and what you are trying to map. If you are creating maps one-on-one with participants, be sure that you are using consistent methods across observation. In most community mapping projects, a facilitator provides participants with a base map, and asks them guiding questions such as “Where is your doctor’s office?” or “Where do people go when they need to relax?” or “What community supports are available to individuals in this neighborhood?” The map is merely a tool to allow you to facilitate reporting on the geographic and cartographic dimensions of what you are studying. The data collected here will help you contextualize your work, and may even help you identify non-geographic details (e.g., other stakeholders).

Step 4: Synthesize, analyze, and interpret your data.

Once you have collected the data, it is useful to compile the data into a single map, or into several maps that relate to different aspects of your evaluation. If possible, you can have participants review your synthesized maps and ask them what they think and if they have anything to add.

One thing to keep in mind when choosing a participatory method, such as community mapping, is to remember that not all mapping is focused on mapping geographic factors. There are a great variety of participatory methods out there that utilize facilitation techniques with similar levels of engagement as community mapping (See **Appendix 1**). These activities focus on mapping processes, ideas, timelines, or other important dimensions related to your work. You will find that many of these share a lot in common with participatory mapping, but it is important to recognize that community mapping, as a methodology, is particularly developed for mapping physical and virtual social geographies.

Structured Conceptualization

According to Trochim & Kane (2005), 'Concept mapping' refers to any methodology that is used to produce a picture or map of the ideas or concepts developed by an individual or group in relation to their subject matter of interest. Concept mapping is sometimes referred to as 'idea mapping', 'mind maps', 'causal mapping', or 'cognitive mapping'. However, while most concept mapping methods are designed to enhance a person's creative thinking, evaluation requires a more structured form of 'concept mapping' that is sometimes described as *Structured Conceptualization*.

Table 8. How to conduct structured conceptualization studies

Step 1: Preparation

The first step is to focus the mapping project by selecting your topic and the participants you will include. You will also need to coordinate the logistics of the event, such as when and where it will take place.

Step 2: Idea Generation

The next step is to begin generating ideas, usually through some sort of brain storming activity that illicit the information of interest. This can be done live or via teleconference.

Step 3: Organizing Ideas

In step 3, ideas are grouped by asking participants to sort ideas into piles. Participants can use as many or as few piles as they would like. Each participant's piles are then analyzed (See **Figure 7a and 7b**) and you count up how frequently each idea was classified with the other ideas developed. This is known as a correlation or co-occurrence matrix, which includes as many rows and columns as there were original ideas.

Step 4: Evaluating importance

In step 4, ideas are rated by participants across one or more variables of interest (e.g., importance, feasibility). The scores can then be visually depicted in a dot plot with one variable on the x axis and another on the y axis (See **Figure 7c**) – thus allowing you to easily identify which ideas are stronger and which ideas are weaker based on your selected criteria.

Step 5: Concept Mapping

In step 5, the ideas that are more similar to one another (as measured by the value of their correlation matrixes) are placed together into groups (See **Figure 7d**). The participants can then come together again and name these larger clusters. The relative importance of each factor can be represented by the average or summed score for the variables of interest (i.e., ideas that had higher ratings across criteria are more prominent).

The purpose of structured conceptualization is to facilitate a more rigorous methodology for aggregating the views of stakeholders on questions of importance that might arise during the preliminary evaluation (e.g., What should the program look like? What is important to evaluate?). Of course, depending on your needs, more traditional brainstorming and concept mapping approaches can also be relied on during your evaluation.

Structured conceptualization is a mixed methods approach that involves five major steps (See **Table 9**) and aims to provide a series of maps that depict composite thinking of a group. I want to note that the quantitative aspects of the protocol may or may not be implementable, depending on the skill level of the evaluation team. Thus, I have not gone into depth on explaining how co-occurrence ratings can actually be used to construct the map and how multivariate methods can be used to assign ideas to groups. In any case, the steps I have provided, should be easily implementable – even if the more advanced mapping techniques are not feasible.

a. Individual pile-sort co-occurrence matrices

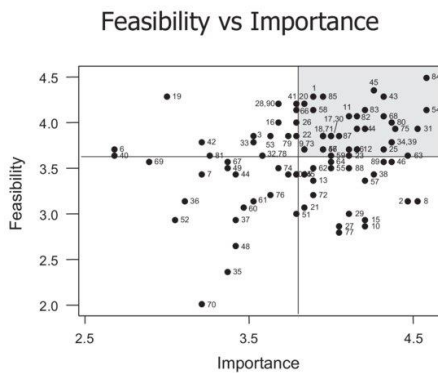
Rater 1	1	2	3	4	5	6	7	8	9	10	11
Idea 1	1	0	0	0	0	0	0	0	0	0	0
Idea 2	0	1	0	0	0	0	0	0	0	0	0
Idea 3	0	0	1	0	0	0	0	0	0	0	0
Idea 4	1	0	0	1	0	0	0	0	0	0	0
Idea 5	0	0	0	0	1	0	0	0	0	0	0
Idea 6	0	1	0	0	0	1	0	0	0	0	0
Idea 7	0	0	0	0	0	0	1	0	0	0	0
Idea 8	0	0	0	0	0	0	0	1	0	0	0
Idea 9	0	1	0	0	0	1	0	0	0	0	0
Idea 10	0	0	0	0	0	0	0	0	1	0	0
Idea 11	0	0	1	0	0	1	0	0	0	0	1

Rater 2	1	2	3	4	5	6	7	8	9	10	11
Idea 1	1	0	0	0	0	0	0	0	0	0	0
Idea 2	0	1	0	0	0	0	0	0	0	0	0
Idea 3	0	0	1	0	0	0	0	0	0	0	0
Idea 4	0	0	0	1	0	0	0	0	0	0	0
Idea 5	0	1	0	0	1	0	0	0	0	0	0
Idea 6	0	1	0	0	0	1	0	0	0	0	0
Idea 7	0	0	0	1	0	0	1	0	0	0	0
Idea 8	0	0	0	0	0	0	0	1	0	0	0
Idea 9	0	1	0	0	0	1	0	0	0	0	0
Idea 10	0	0	0	0	0	0	0	0	1	0	0
Idea 11	0	0	0	0	0	1	0	0	0	0	1

b. Summed co-occurrences matrix

Summed	1	2	3	4	5	6	7	8	9	10	11
Idea 1	1	0	0	0	0	0	0	0	0	0	0
Idea 2	0	1	0	0	0	0	2	0	0	0	0
Idea 3	0	0	1	0	0	0	0	0	0	0	0
Idea 4	1	0	0	1	0	0	0	2	0	1	0
Idea 5	0	1	0	0	1	0	0	0	0	0	0
Idea 6	0	2	0	0	0	0	0	0	0	2	0
Idea 7	0	0	0	1	0	0	0	0	0	0	0
Idea 8	0	0	0	0	0	0	0	0	0	0	0
Idea 9	0	2	0	0	0	2	0	0	0	0	0
Idea 10	0	0	0	0	0	0	0	0	0	0	0
Idea 11	0	0	1	0	0	2	0	0	0	0	1

c. Scatterplot showing feasibility and importance rankings.



d. cluster mapping showing ideas that were scored more closely together.

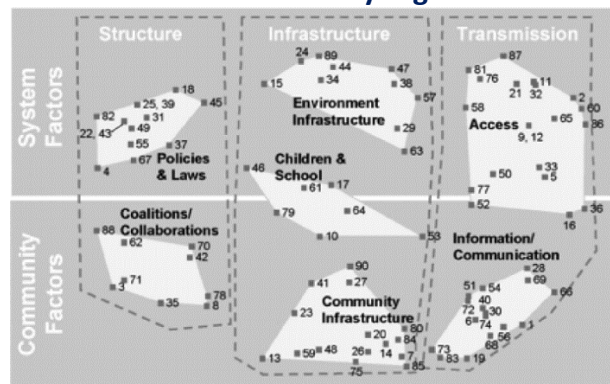


Figure 8. Explanatory Figures for Structured Conceptualization Methodology

Delphi Method

Another important methodology that can be used to illicit the opinions of key informants and experts is known as the **Delphi Method**. The Delphi method is a systematic and interactive way of soliciting opinions from experts. Usually conducted via questionnaires over multiple rounds, the Delphi method aims to provide insight into areas off agreement and disagreement to develop a consensus opinion. Because the survey is administered via questionnaires, it can be used when you want to provide anonymity to the participants.

A Delphi study might follow the following steps (See **Figure 9**):

- A questionnaire is designed.
- Participants are invited to take part in the survey.
- The first questionnaire is administered.
- Responses from the first round are analyzed.
- A second questionnaire is designed based on results of the first.
- The second questionnaire is administered.
- Responses from the second round are analyzed.
- Scores are reported back to participants, who may revise their scores based on the answers of other participants.
- Responses from the third round are analyzed.
- Results are reported.

For example, the first round of the survey might ask experts “Which clinical area is of high priority for developing an intervention?” and provide them with a list of clinical areas (e.g., Cancer, Mental Health, Vaccinations). After analyzing responses, the second round questionnaire might say something like, “Cancer has been identified as a priority clinical area for intervention, what barriers to care are most significant for cancer patients?” Additional surveys scoping deeper and deeper into the issue can be conducted. At some point in the Delphi analysis, participants should be given the opportunity to revise their results after seeing aggregate scores from other participants. For example, if 60% of participants listed cost of accessing care as a leading barrier to care, experts who had not considered cost might recognize that they are underestimating the weight of cost concerns and revise their responses accordingly.

Delphi analysis therefore provide a methodology that can rapidly build consensus, even if participants are able to meet together. Further, while it allows for individuals to make their decisions in collaboration with other experts, individuals are less vulnerable to “group think” (e.g., where group dynamics cause individuals to rely on the opinion of dominant voices) that might inhibit focus groups. As a low cost method, it does allow for relatively large amounts of data to be collected, but it may still nevertheless be time consuming. In the absence of participant motivation to complete the interactive surveys, it may be difficult to keep people engaged – especially compared to one-off interactions such as interviews or focus groups.

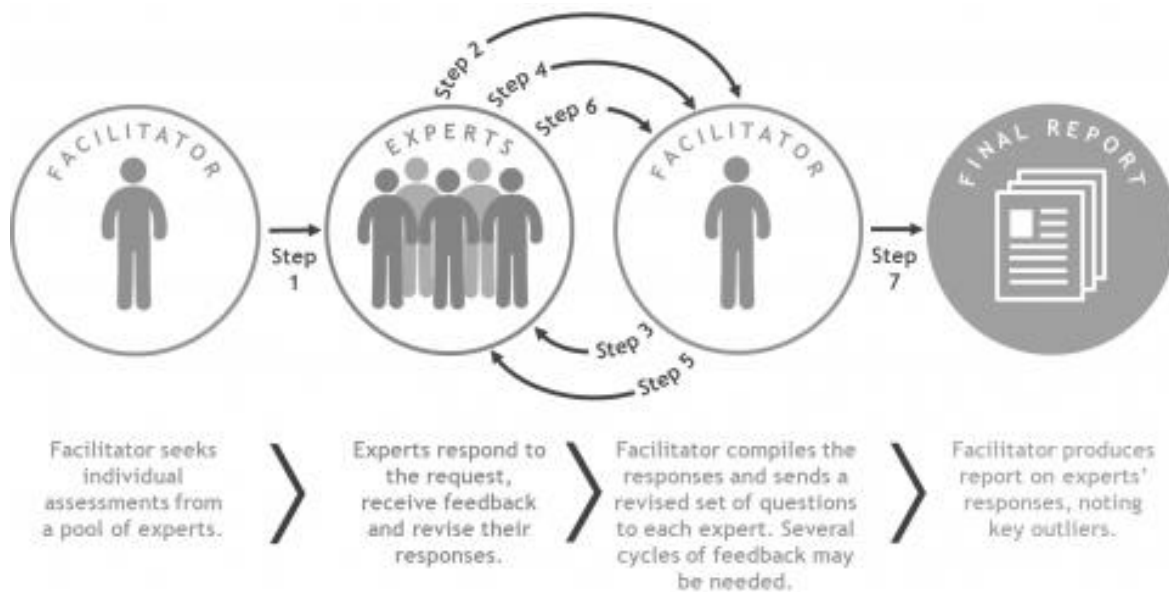


Figure 9. How to use the Delphi method.

One final warning should be considered if you are interested in using the Delphi method. That is, if the panelists invited to participate are not well informed regarding the topic at hand, it is possible that the consensus reached would mislead the evaluation. Moreover, the confirmation of biases through the Delphi method might reinforce wrong views among panelists. If decision makers or other leaders are participating in such a Delphi analysis, you may find that the Delphi method leads these individuals to perpetuate these incorrect viewpoints. This is why it is absolutely vital that expert patients, care providers, policy analysts, and topic experts be included – to ensure that a diverse and highly experienced panel is able to provide the best information available.

Environmental scans (PESTLE)

Understanding the external environment or context of an intervention and its evaluation is a necessity for any successful activity. However, environments are complex and it can be difficult to ascertain what merits your attention and what does not. PESTLE analysis is an environmental scanning technique that provides a framework for you to think through your environmental context in order to help you focus your analysis. In particular, PESTLE reminds you to think through the **Political**, **Economic**, **Social**, **Technological**, **Legal**, and **Environmental** factors that shape the external environment.

Let's take these categories one by one:

- In considering the **political** factors at play you will want to consider how the government will impact your actions. Are there government policies that restrict what you can do? Do tax laws, tariffs, or other regulatory pressures need to be considered? Do you have support of the current government? Is the current government positioned to back your efforts?
- Relatedly, you will want to think through how broad **economic** trends might be working in your favor. Are the monetary and fiscal policies favorable? Is the economy strong? Are unemployment rates stable? Is the program being considered cost-effective?
- For **social** factors, the number of considerations start to multiply exponentially. You will have to think about generational shift, culture, community expectations, social burden, attitudes and tensions in the community.
- For **technological** factors, you need to be thinking through how technological innovations might benefit your intervention, how quickly your intervention will be deemed obsolete, and whether you meet current technical standards with respect to privacy, security, and other standards.
- **Legal** factors must also be considered and include any regulations or liabilities that might impact your program. You will need to ensure that any policies or procedures align with existing law.
- Finally, **environmental** factors can be wide ranging, but can include the geographic location of your communities and health services, the

environmental impact of your program, how climate and other factors might impact program delivery, and so on.

How you address these categories and identify all of the relevant external conditions associated with each is up to you. In some cases, PESTLE might be used to guide a working group through a brainstorming session, in other cases an evaluator might do much of the leg work necessary to understand these factors. However, remember that the focus of the PESTLE analysis is to understand how these external conditions will impact your organization, intervention, or evaluation. Therefore, it is important to revisit each of the PESTLE criteria as your activities evolve and take shape.

SWOT Analysis

PESTLE Analysis can be used hand-in-hand with another analysis technique known as SWOT, which is an acronym for Strengths, Weaknesses, Opportunities, and Threats. The first two factors, strengths and weaknesses, focus on the internal context of an organization. The second two factors, opportunities and threats, relates to the external environments. Therefore, combining SWOT and PESTLE allows you to consider how each of the PESTLE categories might represent opportunities or threats to your organization. On the other hand, the focus on internal context helps you to assess readiness and stability of your own organization by thinking through your current organization strengths and weaknesses. In conducting a SWOT analysis you will want to follow four general steps:

- Collect data that will help you understand your organizations current standing and any forthcoming factors worth considering.
- Organize key facts and figures into the SWOT framework.
- Create SWOT matrices and tailor them to each activity or program you are interested in conducting.
- Consider the SWOT factors relevant to your program and make decisions on how to proceed with consideration of these factors.

The ultimate goal of a SWOT analysis is to help you think through the key factors at play and focus you and your organization's efforts on those activities which are best suited to your current capabilities. Regular SWOT analyses will help your organization adapt to changing circumstances and improve your capacity to meet the needs of those you aim to serve.

Task Analysis

While SWOT and PESTLE analysis can help you identify some of the macro- and meso-level factors that might influence the decisions you make surrounding your work, Task analysis provides you with a tool for micro-level analysis. By micro-level I mean the manual and mental activities that make up a given activity. Every intervention is composed of a variety of activities. Some of these activities are relatively complex, while others are simple. Some require great mental skill, while others can be done routinely. Task analysis gives you a process to think about what is needed for successful completion of a task.

When you think about task analysis, you should be thinking through more than the IKEA-like instructions that outline step by step what is to be done (though this is a part of task analysis) and think through also what tools you should provide and which parts you should provide extras of – in anticipation that at least one will go missing along the way. Indeed, task analysis should be a comprehensive process wherein systems are broken down into operations and operations are broken down into tasks and tasks are broken down into steps, with pre- and post- steps considered as well.

Completing a task analysis will help you consider more thoroughly which steps are necessary, which processes or operations can be merged, who should be assigned with tasks, and how you can prepare individuals to succeed best in their work. Task analysis, as a form of preliminary evaluation, is therefore an exceptionally useful tool to consider prior to planning your process evaluation.

As with all the other methods presented in this text, there is some variability in the steps undertaken as part of a task analysis. The steps listed below provide a general approach used in most task analyses:

- Define the task under investigation and identify the purpose of the task analysis.
 - Be sure to consider related tasks – very few tasks are discreet.
- Observe the task or obtain data about how the task is performed.
- Break down the task into its manual and mental steps.
- Analyze, perhaps through consultations with participants, whether all steps are necessary and which steps might pose a challenge to participants.

- List the revised list of steps associated with each operation and identify any prerequisites for each step. Descriptions of steps should be presented in an appropriate level of detail as is needed for comprehension.
- Sometimes task diagrams are created (See **Figure 10**).

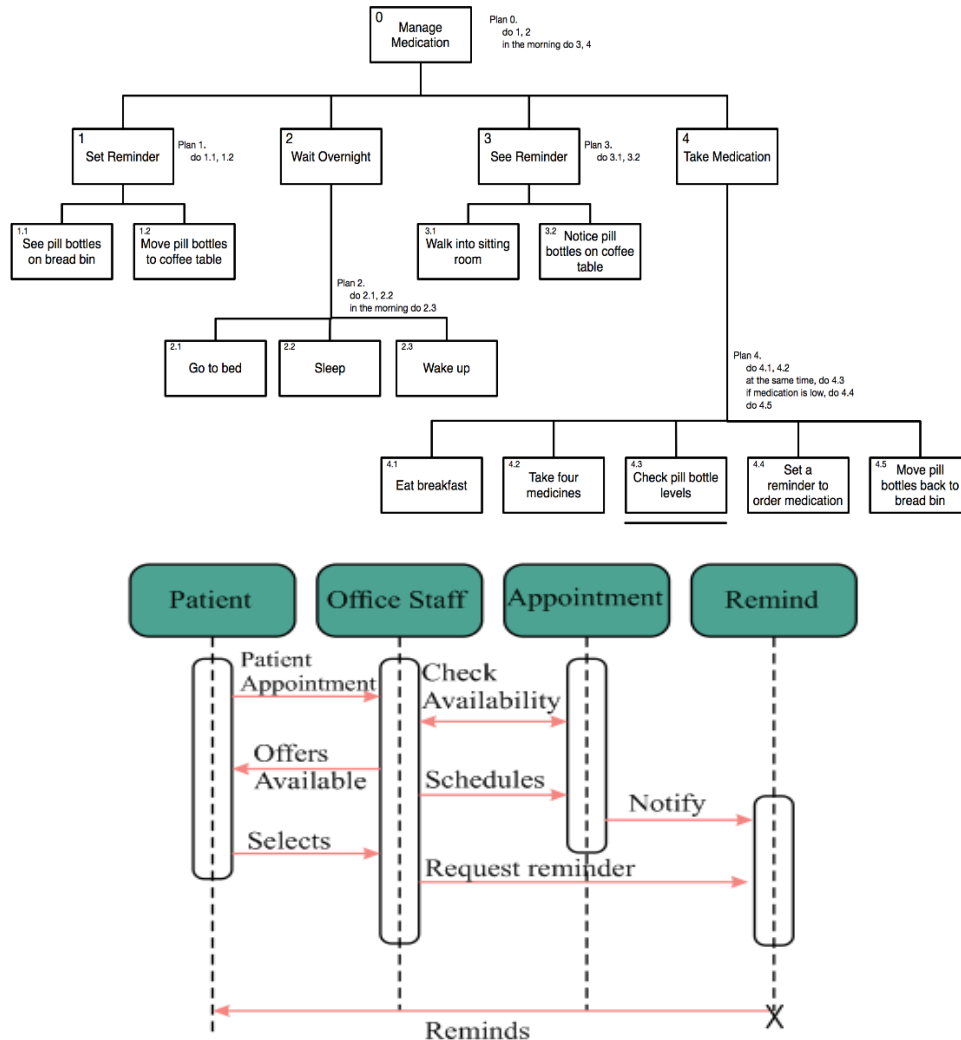


Figure 10. Hierarchical Task Diagram (Top) and Sequential Task Diagram (Bottom)

Cognitive Task Analysis (CTA) is closely related to task analysis. While the methodology can be quite similar, CTA focuses on the cognitive steps associated with a task. This is particularly important for understanding how experts function and complete their work. Indeed, imagine for instance the task of diagnosing an internal illness or conducting a psychotherapy session. These mental tasks are incredibly complex. Yet, it is important that a consistent model be developed if you are to understand these processes.

According to Potworowski et al. (2013) CTA thus tries to uncover

- the step(s) in which the most challenging decisions were made,
- what made those decisions difficult,
- what information was needed to make the critical decisions,
- who needed the information,
- how that information was obtained and transmitted,
- what went wrong or fell through the cracks and why, and
- how the team detected failures and problems in the task.

Typically, this information can be obtained through interviews with those who undertake these tasks. However, when possible, be sure to interview multiple experts so you can get a sense of how much diversity there is in the macro-cognitive schema underlying a given task. Understanding these multiple schemas can help you to identify the necessary and sufficient cognitive steps that underlie a successful cognitive task.

Knowing When to Use the Right Method

At this point, you are probably beginning to feel like our review of the various preliminary evaluation methodologies has become a bit laborious. Therefore, I feel like we should move on, with the acknowledgement that I have not covered every method, nor every version of the methods we have discussed. Before we move to talking about process evaluations, I want to just quickly review the eight preliminary evaluation methods covered in this chapter by quickly articulating when you might want to use them. This is probably the most important part of the chapter. That said, my suggested uses are not definitive and many of the methods can be used for a range of purposes. The information here is merely a guide for you to consider.

As we begin to think about when to use the methods covered, I want to return your attention to the questions we identified as being core to preliminary evaluations. Let's start with "What is being done?" If you want the answer to this question, document reviews, observations, and task analyses are all great systematic methods you could use. If you want to know what you should be doing, try the Delphi method. If you want to know what factors will impact what you will do, structured conceptualization, community mapping, PESTLE, and SWOT analysis all work great. The important thing is that you have a methodology that your stakeholders will have confidence in and that will lay a strong foundation for the work you aim to do with them in the future.

Learning Activities

1. Visit www.solidvictoria.org and conduct a document review of the information available on their website to answer the following questions:
 - a. What steps of the document review process were necessary or unnecessary?
 - b. What is the goal and primary objectives of SOLID OUTREACH?
 - c. What are the primary activities the organization is undertaking?
 - d. Who are the key stakeholders for this organization's activities?
2. Design an observation study relevant to the work of SOLID Outreach and answer the following questions:
 - a. Who would you observe?
 - b. What style of observation would you choose? Why?
 - c. Where would you observe them?
 - d. What indicators might you keep track of as part of your observations?
3. What is a research question that SOLID Outreach might be interested in that would use community mapping as a methodology?
4. Create a task analysis diagram for an activity you complete at least once per week.
5. Design a series of questions you might use at each iteration of a Delphi Method study. Be sure to indicate how a second wave questionnaire might change based on results of the first wave.
6. Conduct a PESTLE or SWOT analysis on a topic of your choosing.
7. Rate each of the following from 0 to 10 based on how important they are as health issues and explain how you chose to rate them the way you did:
 - a. Cardiovascular Disease
 - b. Colon Cancer
 - c. HIV
 - d. Opioid Overdoses

Additional Readings

- [Document Review Example](#)
- [Observations Example](#)
- [Community Mapping Example](#)
- [Structured Conceptualization Example](#)
- [Delphi Method Example](#)
- [PESTLE Analysis Example](#)
- [SWOT Analysis Example](#)
- [Task Analysis Example](#)
- [Cross Jurisdictional Scan Example](#)
- [Literature Review Example](#)

Chapter 3

PROCESS EVALUATIONS

Learning Objectives

- Define what a process evaluation is.
- Be able to plan a process evaluation.



Part 1

What are process evaluations?

Source: www.boomerangedu.com/courses/nursing/

Process Evaluations

In 2018, the Government of British Columbia began (BC) covering emtricitabine / tenofovir disoproxil fumarate as a pre-exposure prophylactic (PrEP) for HIV prevention among HIV-negative men. When taken prior to and after exposure to the HIV virus, PrEP effectively eliminates the ability for the HIV virus to infect the person exposure. Because PrEP is a cost-effective only in high-incidence sexual networks, PrEP is generally restricted to gay, bisexual, and other men who have sex with men who have HIV Incidence Risk Index (HIRI) scores of 10 or greater. The HIRI scale has been widely used, and the cut point of 10 or greater has been hotly contested.

Nevertheless, it is the cut-off value used in BC. Evaluations of BC's use of the HIRI score, and it's cut off, are ongoing. Evaluators will want to examine whether or not those who are declined PrEP due to low cut off scores go on to acquire HIV. Indeed, policy makers are interested in identifying the optimal cut off score that will protect individuals and provide a sustainable standard for other PrEP programs across the globe.

If the evaluation of the PrEP cut-off value is going to be successful, evaluators will need to know details about how the intervention unfolded. For example, they might ask questions such as the following:

- How many people were screened for PrEP?
- How many people received PrEP?
- How many people were denied PrEP due to low HIRI scores?

These sorts of questions aim to understand the process and implementation of PrEP in BC. Depending on what information is collected, and at what level these analyses are performed (e.g., patient charts, summaries of health authorities, distribution data, etc.), evaluators will also need to know whether or not the cut-off score of 10 was honored. In other words, did doctors give PrEP to participants who scored 9 but really wanted to be on PrEP? Did doctors ever deny PrEP to individuals with scores higher than 10? If so, why? These sorts of questions help evaluators understand whether the program had high fidelity, which is a measure of agreement between what was supposed to be implemented and what actually was.

The purpose of process evaluations is to address issues such as those outlined above. These types of evaluations address the Who, What, When, and Where questions about a program, process, or policy. They deal with the implementation of goals and objectives through examining alignment between these and program inputs, activities, and outputs. A list of some common intervention activities that can be evaluated are provided in table 9. Additionally, the process evaluation will help you understand what challenges arose during the implementation process, why participants chose to participate (or not) in the program, and how the program staff feel about the program and the tasks they perform as part of it.

The process evaluation can also help you evaluate the assumptions and theories you used in designing your program. If, and when, a program has been implemented according to plan, its failure indicates that the assumptions or theories underlying the program are incorrect. For example, if few patients with HIRI scores below 10 go on to acquire HIV, the assumption that a HIRI score cut-off of 10 is sufficient would be proven correct, unless it is revealed that those with HIRI scores less than 10 were actually given PrEP. Alternatively, if a large number of people with low HIRI scores do get HIV, this would indicate that the assumptions of the HIRI score are incorrect.

Thus, you can see that a process evaluation is about more than just evaluating inputs, activities, and outputs – it is a fundamental part of the evaluation process that will help you contextualize the results of the outcome and impact evaluations that you also use.

Table 9. Selected Intervention Activities

Implementation Strategies	Definition	Target Level
Conduct educational meetings	Hold meetings involving program targets (e.g., providers, administrators, other organizational stakeholders, and community, patient/consumer, and family stakeholders) to improve knowledge about the ideal practice.	Patient, Provider
Work with educational institutions	Encourage educational institutions to train providers in the ideal practice.	System
Model and simulate change	Have experts/leaders/respected colleagues model or simulate the ideal practice.	Organization, Provider
Prepare patients/consumers to be active participants	Prepare patients/consumers to be active in their care - e.g., to ask questions about the ideal practice, and evidence behind the ideal practice.	Patient, Provider
Alter payments to health workers	Change ways in which providers are paid for providing the ideal practice.	System
Change record systems	Change records systems to allow better capturing of patient information and assessment of implementation or clinical outcomes related to the ideal practice; for example electronic patient records, or systems for recalling patients for follow-up or prevention e.g., immunization.	System, Organization
Change service sites	Change the setting where the ideal practice is provided; for e.g., home vs. healthcare facility, inpatient vs outpatient, specialized vs. non specialized facility, walk in clinics, medical day hospital, mobile units.	System, Organization
Alter incentive/ allowance structures	Work to incentivize or disincentivize the adoption and implementation of the ideal practice.	System, Organization
Change accreditation or membership requirements	Strive to alter accreditation standards so that they require or encourage use of the ideal practice. Work to alter membership organization requirements so that those who want to affiliate with the organization are encouraged or required to use the ideal practice.	System, Organization
Conduct local consensus discussions	Engage local providers and other stakeholders in discussions about whether the chosen problem is important and whether the selected practice to address it is appropriate; e.g., agreeing on a clinical protocol to manage a patient group or adapting a guideline for a local health system.	Organization, Provider, Patient
Audit and provide feedback	Collect and summarize performance data related to the ideal practice over a specified time period and give it to providers and administrators to monitor, evaluate, and modify behavior.	Provider
Create or change credentialing and/or licensure standards	Create or change credentialing and/or licensure standards related to the ideal practice.	System
Develop/alter scope of practice standards	Develop evidence-based policies that regulate what health professionals are able to do in their role, or alter existing scope of practice standards to include the ideal practice.	System

Process Evaluation Frameworks

While there are a variety of evaluation frameworks out there, two in particular stand out to me as being particularly useful: The Consolidated Framework for Implementation Research (CFIR) and the RE-AIM framework. Neither framework exclusively fits with-in the PPOI (preliminary-process-outcome-impact) framework for evaluation that this text uses, but I still want to introduce these to you. The reason I include them in this section is that both models have a strong emphasis on implementation and process. Further, both of these two frameworks are used quite extensively and have great help-documentation websites:

- <https://cfirguide.org/>
- <http://www.re-aim.org>

So, let me briefly introduce you to RE-AIM and then I will discuss CFIR.

RE-AIM

RE-AIM stands for Reach, Effectiveness, Adoption, Implementation, and Maintenance.

- ***Reach*** measures the number, proportion, or representativeness of individuals who were recipients of the ideal practice. Reach Measures may include: the number of potential participants approached, the number of participants deemed ineligible to participate, the number of participants that actually participated, the percent of all eligible invited participants who accepted participation, and the characteristics of participants compared with nonparticipants.
- ***Effectiveness*** measures the impact of an evidence-based practice on outcomes. It asks questions that are similar to those that will be described later in the sections on outcome and impact evaluations, including whether the intervention made a difference, how much of a difference it made, and whether some program recipients benefited more than others.
- ***Adoption*** measures the number, proportion, or representativeness of settings/sites that decided to implement an intervention. Measures for adoption would include things like the number of settings in a given population qualified to host the intervention, the number of settings that were interested in participating, the number of settings that were not appropriate for the study, the number of settings that met criteria and chose to participate, the percent of the

total number of available settings that actually participate, and the characteristics of participating settings compared with nonparticipating settings. At an individual level, adoption could focus on specific Healthcare providers and physicians, or even on consumers of an intervention. The basic question is how many of those reached participated.

- **Implementation** measures how well the strategies were implemented (i.e., implementation quality). It includes measures of dose (e.g., how much of an intervention did someone complete), adherence/fidelity (e.g., how closely did the implementation adhere to the planned program), adaptation (e.g., how well was the program adapted for its environment), and quality of delivery (e.g., what qualities of the program could be improved or what worked well?).
- **Maintenance** focuses on three elements: sustainability, scale-up, and spread.
 - **Sustainability** is defined as a program or implementation strategy that continues to be delivered after a defined period of time and leads to changes in behavior that are aligned with an ideal practice leading to continued production of benefits for individuals and systems. Planning for sustainability makes it more likely that you will sustain the intervention. Not sustaining the implementation intervention decreases the chances that the intervention will be adopted in the future. Unfortunately the literature on sustainability is scant and it is difficult to sustain many programs given current funding models and changes to funding priorities.
 - **Spread**, also called horizontal diffusion, is the process by which an innovation, best practice, or knowledge product is communicated to other Healthcare settings. If I were to adopt a youth-based intervention developed in the U.S. – I would be dealing with spread.
 - **Scale-up**, also called vertical diffusion, is the process of increasing coverage, range, or sustainability of services.

CFIR

CFIR was developed in 2008, and includes consolidated frameworks from many health evaluation theories based on a systematic review of over 500 published sources across 13 scientific disciplines. CFIR is considerably more complex than the RE-AIM framework, but it broadly focuses on five main constructs: (1) the characteristics of an intervention, (2) the external environment or outer settings, (3) the internal environment or inner

setting, (4) the characteristics of individuals participating in the study, and (5) the process by which an intervention is implemented. **Table 10** reviews these constructs in greater detail.

Table 10. Consolidated Framework for Implementation Science

Construct	Short Description
I. INTERVENTION CHARACTERISTICS	
A Intervention Source	Perception of key stakeholders about whether the intervention is externally or internally developed.
B Evidence Strength and Quality	Stakeholders' perceptions of the quality and validity of evidence supporting the belief that the intervention will have desired outcomes.
C Relative Advantage	Stakeholders' perception of the advantage of implementing the intervention versus an alternative solution.
D Adaptability	The degree to which an intervention can be adapted, tailored, refined, or reinvented to meet local needs.
E Trialability	The ability to test the intervention on a small scale in the organization, and to be able to reverse course (undo implementation) if warranted.
F Complexity	Perceived difficulty of the intervention, reflected by duration, scope, radicalness, disruptiveness, centrality, and intricacy and number of steps required to implement.
G Design Quality and Packaging	Perceived excellence in how the intervention is bundled, presented, and assembled.
H Cost	Costs of the intervention and costs associated with implementing the intervention including investment, supply, and opportunity costs.
II. OUTER SETTING	
A Patient Needs and Resources	The extent to which patient needs, as well as barriers and facilitators to meet those needs, are accurately known and prioritized by the organization.
B Cosmopolitanism	The degree to which an organization is networked with other external organizations.
C Peer Pressure	Mimetic or competitive pressure to implement an intervention; typically because most or other key peer or competing organizations have already implemented or are in a bid for a competitive edge.
D External Policies and Incentives	A broad construct that includes external strategies to spread interventions, including policy and regulations (governmental or other central entity), external mandates, recommendations and guidelines, pay-for-performance, collaboratives, and public or benchmark reporting.
III. INNER SETTING	
A Structural Characteristics	The social architecture, age, maturity, and size of an organization.
B Networks and Communications	The nature and quality of webs of social networks and the nature and quality of formal and informal communications within an organization.
C Culture	Norms, values, and basic assumptions of a given organization.
D Implementation Climate	The absorptive capacity for change, shared receptivity of involved individuals to an intervention, and the extent to which use of that intervention will be rewarded, supported, and expected within their organization.
1 <i>Tension for Change</i>	<i>The degree to which stakeholders perceive the current situation as intolerable or needing change.</i>
2 <i>Compatibility</i>	<i>The degree of tangible fit between meaning and values attached to the intervention by involved individuals, how those align with individuals' own norms, values, and perceived risks and needs, and how the intervention fits with existing workflows and systems.</i>
3 <i>Relative Priority</i>	<i>Individuals' shared perception of the importance of the implementation within the organization.</i>

4	<i>Organizational Incentives and Rewards</i>	<i>Extrinsic incentives such as goal-sharing awards, performance reviews, promotions, and raises in salary, and less tangible incentives such as increased stature or respect.</i>
5	<i>Goals and Feedback</i>	<i>The degree to which goals are clearly communicated, acted upon, and fed back to staff, and alignment of that feedback with goals.</i>
6	<i>Learning Climate</i>	<i>A climate in which: a) leaders express their own fallibility and need for team members' assistance and input; b) team members feel that they are essential, valued, and knowledgeable partners in the change process; c) individuals feel psychologically safe to try new methods; and d) there is sufficient time and space for reflective thinking and evaluation.</i>
E	Readiness for Implementation	Tangible and immediate indicators of organizational commitment to its decision to implement an intervention.
1	<i>Leadership Engagement</i>	<i>Commitment, involvement, and accountability of leaders and managers with the implementation.</i>
2	<i>Available Resources</i>	<i>The level of resources dedicated for implementation and on-going operations, including money, training, education, physical space, and time.</i>
3	<i>Access to Knowledge and Information</i>	<i>Ease of access to digestible information and knowledge about the intervention and how to incorporate it into work tasks.</i>
IV. CHARACTERISTICS OF INDIVIDUALS		
A	Knowledge and Beliefs about the Intervention	Individuals' attitudes toward and value placed on the intervention as well as familiarity with facts, truths, and principles related to the intervention.
B	Self-efficacy	Individual belief in their own capabilities to execute courses of action to achieve implementation goals.
C	Individual Stage of Change	Characterization of the phase an individual is in, as he or she progresses toward skilled, enthusiastic, and sustained use of the intervention.
D	Individual Identification with Organization	A broad construct related to how individuals perceive the organization, and their relationship and degree of commitment with that organization.
E	Other Personal Attributes	A broad construct to include other personal traits such as tolerance of ambiguity, intellectual ability, motivation, values, competence, capacity, and learning style.
V. PROCESS		
A	Planning	The degree to which a scheme or method of behavior and tasks for implementing an intervention are developed in advance, and the quality of those schemes or methods.
B	Engaging	Attracting and involving appropriate individuals in the implementation and use of the intervention through a combined strategy of social marketing, education, role modeling, training, and other similar activities.
1	<i>Opinion Leaders</i>	<i>Individuals in an organization who have formal or informal influence on the attitudes and beliefs of their colleagues with respect to implementing the intervention.</i>
2	<i>Implementation Leaders</i>	<i>Individuals from within the organization who have been formally appointed with responsibility for implementing an intervention as coordinator, project manager, team leader, or other similar role.</i>
3	<i>Champions</i>	<i>"Individuals who dedicate themselves to supporting, marketing, and 'driving through' an [implementation]" [101] (p. 182), overcoming indifference or resistance that the intervention may provoke in an organization.</i>
4	<i>External Change Agents</i>	<i>Individuals who are affiliated with an outside entity who formally influence or facilitate intervention decisions in a desirable direction.</i>
C	Executing	Carrying out or accomplishing the implementation according to plan.
D	Reflecting and Evaluating	Quantitative and qualitative feedback about the progress and quality of implementation accompanied with regular personal and team debriefing about progress and experience.

Learning Activities

1. Find an example of a process evaluation from searching a literature database (e.g., Google Scholar, PubMed) and answer the following questions:
 - a. Who was involved in the process evaluation?
 - b. What did those involved in the process evaluation do?
 - c. What questions did the process evaluation aim to address?
 - d. What indicators did they measure as part of their process evaluation?
2. Imagine you are asked to conduct a process evaluation of a dietary intervention by the United Kingdom's National Obesity Observatory. Create a list of indicators you would want to collect to measure the (a) inputs, (b) activities, and (c) outputs associated with the dietary intervention.
3. In thinking about the evaluator-stakeholder relationship, what tensions do you think might be introduced by conducting a process evaluation? What do you think you can do to help manage these tensions and maintain strong relationships with the program staff whose work you are evaluating?
4. Patient-uptake and access to the program is an important factor to consider when conducting a process evaluation. If you did not already identify indicators related to patient uptake, what sort of data would you collect to understand this aspect of the program's development?
5. Visit the RE-AIM and CFIR websites and create a few questions that you would measure for each part of each framework.

Additional Readings

- Moore et al. (2015). "Process evaluation of complex interventions: Medical Research Council guidance." *BMJ*.
- Schneider et al. (2009). "Rationale, design and methods for process evaluation in the HEALTHY study." *International Journal of Obesity*.
- Anderson et al. (2014). "Process Evaluation of Workplace Interventions with Physical Exercise to Reduce Musculoskeletal Disorders." *International Journal of Rheumatology*.
- Williams et al. (2007). "Process evaluation of a nurse-led telemonitoring programme for patients with asthma." *Journal of Telemedicine and Telecare*.

Part 2

How do I plan a process evaluation?



Source: <http://www.boomerangedu.com/courses/nursing/>

Selecting Indicators and Data Sources

From the last section, you can easily get a sense of the many things can be measured in the course of a process evaluation. However, it is not feasible, or even necessary, to measure every detail of a program during a process evaluation. While frameworks such as CFIR and RE-AIM will help guide you in what you should measure, process indicators should always be informed by the priorities of the stakeholders and evaluation sponsors and emphasize the primary elements developed in the logic model. While some elements of the logic model may receive more attention than others, in general a few indicators for each component are all that is needed.

In general, true and thorough process evaluation indicators should capture:

- inputs (*e.g., were the resources available as intended*),
- activities (*e.g., were the activities delivered as planned*),
- outputs (*e.g., did the project meet its targets*),
- mechanisms of action (*e.g., how the inputs impact the planned activities and how the planned activities impact the outputs*),
- controls (*e.g., what limitations or constraints impacted the program*), and
- timing (*e.g., the sequencing of activities*).

These six broad domains of process indicators are also sometimes categorized into one of two categories:

- **leading indicators** measure variables that *lead* to results.
- **lagging indicators** measure variables that *follow* a program's implementation.

To understand how these two types of indicators are used in evaluation, consider standard evaluation indicators used in occupational health and safety. Probably the first that comes to mind is "Number of Days without an Accident." This is a lagging indicator. It is a result that emerges from successful implementation of resource inputs and activities that prevent workplace accidents. Other lagging indicators might include the rate of occupational accidents (i.e., Accidents / 1000 employees), the number of sickness absence days, the severity rate of occupational accidents, the number of safety violations, the cost of occupational accidents, and the number of workers reporting work-related disease.

Often times, the leading indicators get less attention. Remember, leading indicators focus on inputs and activities. They lead to the results that are measured by lagging indicators. Example indicators in workplace safety evaluations might include the number of workers who participate in workplace safety training, the number of occupational risk assessments conducted, the number of safety risks remediated, or the number of wellness activities undertaken by employees (assuming wellness activities reduce risk by reducing stresses which might distract employees from following safety protocols). It's worth noting that according to research by Pawlowska ([2015](#)) these sorts of leading indicators tend to be the ones that set high performing companies apart from those with poorer safety records – suggesting that a focus across all indicators provides a better evaluation framework for improving their occupational safety programs.

In many cases, as in the occupational health and safety sector, indicators will be widely available and process evaluations will follow a somewhat standard protocol. It is always worth your while to search out existing indicators and to consider their use in your evaluation. However, often times you will also need to develop indicators specific to your project. If this is the case the following steps can be used to guide the development of these indicators:

- Consult existing literature for relevant indicators.
- Ask stakeholders for input on indicators and data sources.

- Review evaluation questions and use the logic model to identify indicators that will link inputs to activities, activities to outputs, and outputs to outcomes.
- Review each indicator and ensure that it is specific, observable, and measurable.
- If you are trying to measure change in an indicator, establish at which time points you need to assess them at (e.g., baseline, midpoint, end of study).
- Consult stakeholders and program sponsors to ensure that the indicators provide useful information, are feasible to collect or observe, and meet the needs of the evaluation.

An indicator matrix can be used to capture the key information about each indicator and is a useful tool to accompany your logic model. Indicator matrices often include the following basic information for each goal, objective, input, activity, output, outcome, and impact indicator you are interested in measuring:

- Indicator (e.g., what are you measuring?)
- Definition (e.g., how is it calculated?)
- Baseline (e.g., what is the current value?)
- Target (e.g., what value will indicate success of the program?)
- Data source (e.g., how will it be measured?)
- Frequency (e.g., when will you measure it?)
- Accountability (e.g., who will be responsible to measure it?)
- Reporting (e.g., who will it be reported to?)

Temporal Considerations

One of the things you will have to consider when selecting indicators is the role that time will play in the analysis. How much time do you have to conduct the evaluation? At what stage is the project already in? What indicators do we expect to be measurable at the current point in time? These are the sorts of questions you should be thinking about. One tool that is widely used in program management and evaluation strategies is the Gantt chart, which is named after its inventor, Henry Gantt. Gantt charts illustrate which projects will take place at which points in time. To create a Gantt chart, you should begin by listing all of the project activities. Within a process evaluation – these are the things to be evaluation. You should list program activities from first to last, with those which are dependent on earlier activities listed after the activities they depend on. Generally, each activity is listed in its own row. Columns are then made to provide the

time scale of each activity – usually at the weekly, monthly, or quarterly level. I provide an example format made in excel.

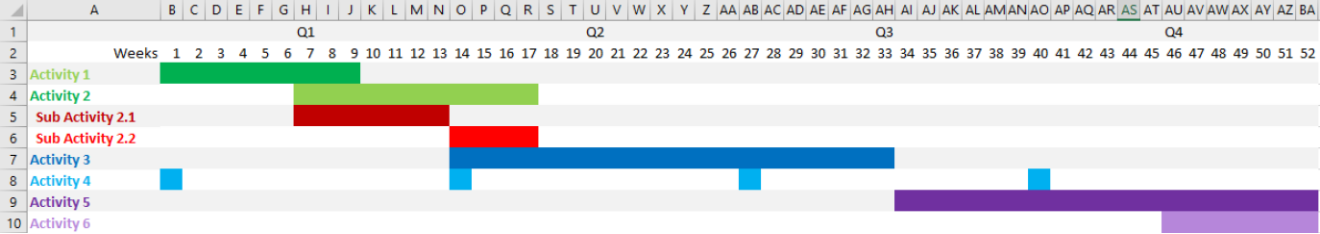


Figure 11. Format of Gantt chart

Use Gantt charts to help you think about how a project is to be organized; and incorporate evaluation activities into your Gantt chart to make it clear at what stage in the course of a project each evaluation activity will be undertaken – this will help ensure that your evaluation is directly connected with program activities.

Theory of Change

A final piece of the puzzle needed to develop good process indicators is a connection to the theoretical framework that underlies a specific program or activity. Earlier in this text we talked about how theory-based interventions have an explicit focus on evaluating the theory about how something works. This will often be the case with interventions that you might be asked to evaluate. It is rare that there is not some rationale guiding the work of an organization – even if the rationale is not formally articulated as a theory. For example, if you are doing a social media intervention, the theory is that you can reach more people on social media and that peers sharing information’s with other peers has the potential to influence the outcomes of interest. While this thinking might not be articulated as a formal social theory, you will need to make sure that the logical flow of a process is measured. For instance, you might look at whether paid social media advertisements were engaged with more or less frequently than those shared by individuals. This would allow you to better understand why a part of your program is or is not working or how it could be working better. In many cases you will find that you will need to do some formative work just to understand what the theory underlying a project is and sometimes you might be asked to connect the lay theory that guides the project to a more formal theory from the academic literature. However, this is often an optional step in your evaluation and you should take the approach that you and your stakeholders think will work best.

Learning Activities

1. Select a topic and search an academic or gray literature database (PubMed, Google, etc.) to see if you can identify indicators used in that area. Which of the indicators you identified measure the process of implementation? Which of the process indicators do you like best; and why? What do you think makes a good indicator based on your review of existing indicators?
2. Create a Gantt chart outlining the activities that would be required to implement a program aiming to distribute antiretroviral drugs to patients living with HIV in South Africa.
3. Create an indicator matrix for the process indicators that you would measure for an implementation of a workplace wellness program that is promoting health eating and exercise through posters, information sessions, and flexible work arrangements. Feel free to take some liberties in describing the program you are developing the matrix for.
4. Identify which of the indicators from your newly created indicator matrix are leading indicators and which are lagging indicators.
5. Create a checklist or set of questions that you would ask a stakeholder as you work to develop process indicators for their intervention. Why are these the steps or questions that you think are most important?

Additional Readings

- Brousselle & Buregeya. (2018). “Theory-based Evaluations: Framing the existence of a new theory in evaluation and the rise of the 5th Generation.”
- Brown (2009). “Good Practice Guidelines for Indicator Development and Reporting.” Statistics New Zealand.
- Chimbindi et al. (2018). “Translating DREAMS into practice: Early lessons from implementation in six settings.”

Chapter 4

OUTCOME EVALUATIONS

Learning Objectives

- Define what an outcome evaluation is.
- Be able to plan an outcome evaluation.



Part 1

What are outcome evaluations?

Source: <https://www.cbc.ca/news/canada/toronto/toronto-injection-sites-1.4144096>

Outputs, Outcomes, and Impacts – Oh my!

If a process evaluation aims to measure the inputs, activities, and outputs of an intervention, the outcome evaluation will focus on the outcomes and impacts. I find one of the most difficult things to conceptualize when thinking about the connection between logic models and evaluation strategies is to differentiate between outputs, outcomes, and impacts. We covered this earlier, but it is worth reviewing again here. In doing so, we should recognize that this is just jargon of the field – but nevertheless, jargon is important to learn if your work is to be taken seriously.

So, let's take the example of a soup kitchen. You can imagine that one of the activities a soup kitchen might undertake is providing meals. In this case, providing meals is an output. It would probably be measured by the number of meals served by the soup kitchen in a given month. However, the ultimate goal of soup kitchens, like restaurants, is not to simply serve a meal. Serving meals is merely a means to achieving an outcome. So, what would an outcome of a soup kitchen be? Well, it could be the extent to which the individuals being served by the soup kitchen had their hunger needs met. To evaluate the outcome – reduced hunger among individuals being served – you would probably need to measure more than just what your organization did. Maybe you could administer a hunger scale or food insecurity scale to individuals before and after the

intervention was introduced? Improvements on a food insecurity scale might indicate that the program is contributing to achieving outcomes. Then again, how do we know that improvements on the food insecurity scale are explicitly attributable to the food kitchen? And how do we know if the observed results are actually making a difference? What if scores improved simply because the economy is doing better or because a new school lunch program reduced hunger scores among a segment of your target population? For instance, if you recall the Somerville Youth Study discussed at the outset of this text, it is not entirely clear why those who were involved in the program fared worse off than those who were not involved. It is possible that the effect was due to the group aspect of the summer camps, but there are many other potential explanations that might exist.

This is where impacts come in. Impacts differ from outcomes in that they aim to understand to what extent the outcomes are a result of the program activities and the extent to which the goal of the project is addressed. Impacts are also often considered separate from outcomes in terms of the indicators being used. Food security scales might make a good indicator for measuring outcomes, but it may not be well suited for measuring impacts.

To drive this point home, let's take another example: weight loss. For weight loss, the chief output will be the amount of calories you consume. You might also be interested in the number of calories burned or the minutes of strenuous activity done in the past week. The outcome in this example would be your observed weight; but weight is probably not a good indicator for the impact. In this example, the impact would be the degree to which weight loss improves your health or self-image or whatever your ultimate goal was.

Outcome Evaluations

Outcome evaluations tend to focus on outcomes and impacts. Some authors might also consider outputs as part of an outcome evaluation. I tend to think that outputs are more properly situated in a process evaluation – and I think this is probably the dominant opinion in the evaluation field.

Because outcome evaluations focus on outcomes and impacts, they tend to be a bit more difficult to conduct in a rigorous way. Indeed, while process evaluations are not complication free (e.g., dealing with stakeholders who might be resistant to their

activities being evaluated), it is usually possible to design evaluation strategies that can very easily identify what inputs, activities, and outputs were and were not completed. It may not be as easy measuring outcomes. For instance, participants may be lost to follow-up, meaning you can't collect data from them. As such, it would be difficult to assess how the intervention impacted them. Or, considering the weight loss example, maybe simply counting calories is enough to promote weight loss and the other strategies you undertook were not actually what contributed to the change in your health and weight. In earlier chapters, we will discuss the quantitative methods that can be used to overcome some of these challenges, but suffice it to say here that outcome and impact evaluations should be considered carefully if you are to successfully complete one.

Rationale for Outcome Evaluations

Given the many challenges encountered in planning an evaluation, many stakeholders may want to avoid doing one. However, it is your job to help them understand the benefits of outcome evaluations. One of the tensions you will be faced with is the epistemological perspective of your stakeholders. Broadly speaking Epistemology is a theory about what knowledge is and how it is obtained. Two broad epistemologies are widely influential in public health:

- **Constructivism**, also called Interpretivism or Hermeneutics, is a view in philosophy according to which all "knowledge is a compilation of human-made constructions. Constructivism is closely related to Idealism, which is a view that knowledge is primarily acquired by a priori or innate processes and not derived from experience. A priori knowledge is knowledge that is known independently of experience (that is, it is non-empirical, or arrived at beforehand, usually by reason). It will henceforth be acquired through anything that is independent from experience.
- **Empiricism** is a theory that states that knowledge comes only or primarily from sensory experience. It is closely associated with positivism, which is a philosophical theory stating that certain ("positive") knowledge is based on natural phenomena and their properties and relations. Thus, information derived from sensory experience, interpreted through reason and logic, forms the exclusive source of all certain knowledge. Positivism holds that valid knowledge (certitude or truth) is found only in this a posteriori knowledge. A posteriori

knowledge is knowledge that is known by experience (that is, it is empirical, or arrived at afterward).

Most health workers tend to support empirical work or at least understand its utility, even if they do not strongly agree with its philosophical foundations. As such, outcome evaluations tend to lean towards empiricism. For those, who are resistant to these approaches, you can emphasize the benefits of an empirical approach. Some benefits of an empirical outcome evaluation include:

- **Improved Accountability.** Conducting outcome evaluations ensures you are developing services and programs that work.
- **Improved Services.** If you do outcome evaluations, you will be able to identify problems with your services and strategies on how to improve them.
- **Improved Revenues.** Demonstrating your success will allow you to demonstrate your success to funders, stakeholders, decision makers, and your communities. This means your organization will reap the rewards that come from being an “evidence-based” program.
- **Improved Morale.** Demonstrating the effectiveness of programs can help you show that you are having an impact. This can boost morale and help recruitment and retention – both of program participants and staff.

In a United Way survey on the use of outcomes evaluations, 298 organizations reported that outcome measurements helped them in the following ways:

- Communicate program results to stakeholders and staff (88%)
- Clarify the intended purposes of the program (86%)
- Identify effective practices within the program (84%)
- Successfully compete for resources/funding (83%)
- Enhance record-keeping systems (80%)
- Improve the service delivery of the program (76%)
- Share effective practices with other agencies (72%)
- Inform program participants about outcome measurement results (70%)
- Assess staff performance; identify staff training needs (64%)
- Allocate resources within the program and the agency (61%)
- Increase program participants’ investment in achieving program outcomes (55%)
- Recruit staff or volunteers to work in the program (42%)

Learning Activities

1. Classify whether each of the following is an output, outcome, or impact:
 - a. HIV incidence decreased at the population level following the introduction of pre-exposure prophylaxis drugs that prevent HIV acquisition.
 - b. 10% of Apple employees participated in a yogic stress-reduction class at Apple Headquarters.
 - c. 33/50 participants in a program stopped smoking.
 - d. 84,483 people visited a nutrition website last month.
 - e. People who saw an anti-stigma advertisement on twitter were three times less likely to post something stigmatizing than those who did not see the advertisement.
 - f. 95% fewer cases of diarrhea were identified in a town after introduction of chlorinated water into the municipal water system.
 - g. Cases of scurvy worldwide have declined after citrus fruit was identified and promoted as an important source of ascorbic acid that prevents scurvy.
2. Discuss whether you think a priori or a posteriori evidence is more useful to guiding decisions around public health programs. Support your position with the type of evidence you think is most appropriately matched to your epistemology.
3. Compose an email to a stakeholder who is resistant to the idea of doing an outcome evaluation. Assume this stakeholder has a constructivist epistemology. You can make up details of the evaluation as needed, but try to convey the benefits of doing an outcome evaluation while demonstrating respect for the stakeholder's likely concerns.

Additional Readings

- Birdthistle et al. (2018). "Evaluating the impact of the DREAMS partnership to reduce HIV incidence among adolescent girls and young women in four settings: a study protocol."
- Wilson. (2000). "The myth of objectivity: is medicine moving towards a social constructivist medical paradigm?"

Part 2

What should I measure in an outcome evaluation?



Source: <https://crosscut.com/2019/02/what-does-federal-lawsuit-mean-seattles-safe-injection-sites>

Choosing What to Measure

With the benefits of outcome evaluations in mind, there is a growing recognition that what gets classified in empirical studies as “evidence” is not always sufficient to ensure the effective delivery of healthcare services. Among the things you must consider when planning an outcome evaluation, you should think about whether what is being measured is the *right* outcome. As an evaluator you need to be careful to ensure that the criteria against which a program is judged are sufficient as to justify your final judgements about the program later on down the road.

Take for instance the issue of substance use decriminalization: Decriminalizing drugs may very well lead to increased use among some individuals, but it can also lead to reduced stigma and harms for people who use drugs. If you measure only increased use, but not the potential benefits of reducing stigma, your evaluation will lead to biased conclusions about whether or not decriminalizing drugs is a good policy. Considering the benefits of decriminalization doesn’t mean that you would necessarily come out in favor of decriminalizing drugs – it simply means that you have taken them into account when

making your decision. In a similar way, when planning an outcome evaluation, you will want to consider both the benefits and harms that result from a specific action.

Two prominent strategies that can be applied to outcome evaluations to assess the overall worth of a program are Monetary Evaluations and Multicriteria Analysis (MCA). Let's look at these one by one – starting with Monetary Evaluations.

Financial Analysis

Monetary evaluations are one of the most widely used, and widely respected, methods for evaluating programs and policies. The claim of “cost-effective,” which might be granted at the end of a monetary analysis, is as good as gold when it comes to winning grant funding or provincial sponsorship for a specific program or policy. However, monetary evaluations are a single criterion approach to analysis. They really only care about the monetary elements of an intervention (e.g., how much time, money and resources will need to be dedicated and how much time, money, or resources will be thusly saved) or things that can be “show-horned” into a monetary framework. There are three main Monetary Evaluations used in outcome evaluation: Financial analysis, cost-effectiveness analysis, and Cost-benefit analysis.

Financial Analysis

Financial analysis is an impact assessment that describes the organizations own financial costs and revenues. As such, financial analyses are pretty straightforward: they largely aim to understand how much a particular program costs the organization. Outside costs, sometimes referred to as externalities, are not captured in financial analyses. Neither are the benefits of a program. In Financial Analysis, the financial impacts of a multi-year intervention are discounted and presented in current value. This is true of most multi-year monetary analyses. Discounting is the process of reducing costs based on preferences for benefits now as opposed to later. As this text is not an economics text, we will forgo any protracted discussion on discounting and adjustment. Suffice it to say here, that the costs are not simply subtracted from the value gains to arrive at a net value. However, even these rudimentary calculations can be sufficient.

Cost-effectiveness Analysis

Because program implementers and policy makers are often concerned with factors beyond just the financial impacts to an organization, cost-effectiveness analyses are

widely used to determine whether an intervention or policy implementation is *worth it* – given the costs. For example, if a clinic finds that a one-on-one substance use counselling program results in 17 fewer overdoses in a month and costs \$3,500 to operate; and a group counselling program results in 6 fewer overdoses in a month but costs only \$700 to operate, one could conclude that the cost per prevented overdose under the one-on-one scheme is \$206 CAD, while the cost per prevented overdose under the group scheme is \$117 CAD. In this scenario, the group program would be considered more cost-effective because the per-unit cost of overdose prevention is cheaper. That said, the method misses the reality that 11 additional overdoses occurred.

Cost-benefit analysis

Cost-benefit analysis takes cost-effectiveness analysis one step further by trying to monetize the value of the outcome. Monetization is the process of assigning a dollar value to either a tangible or intangible thing. For instance, the value of a statistical life is used to calculate the cost or savings of each mortality. While there is no agreed upon value for a statistical life, a dollar value can nevertheless be agreed upon for the sake of a given analysis – allowing you to calculate the implied cost of averting a fatality (ICAF) for a given intervention. Various values have been recommended ranging from \$50,000 to \$150,000 USD per quality-adjusted life year to a standard value of between \$8,000,000 and \$10,000,000 USD per life. In a similar fashion, monetary values – even if somewhat arbitrary – can be assigned to many different elements of a project to try and capture the full costs and benefits that result from a program. So, taking our example from earlier, the cost of the 11 extra overdoses can be quantified if you are conducting a cost-benefit analysis. Regardless of what value is ultimately chosen for the cost of an overdose, it is very likely that those extra 11 overdoses would sway your decision given that the cost difference of the two is so small.

While cost-benefit analyses can range in complexity, they do tend to follow five general steps:

- First, the problem and various solutions are identified.
- Second, the cost of each solution is calculated, with attention to not only the cost of direct inputs but also the external costs and opportunity costs associated with each option.

- Third, the benefits are also quantified. Again, attention should be paid to both intended benefits and the unintended benefits.
- Fourth, the net value is calculated and discounted.
- Fifth, sensitivity analyses are performed by adjusting the assumed or implied costs of various costs and benefits so that you can accurately report on the effects of your assumptions.

It is important to recognize that a cost-benefit analysis is usually based on the data of a pilot study or after the completion of an implementation project. Therefore, it is important for you to think about the indicators you will need to calculate the costs and benefits well before your study is concluded. As you will need to make many assumptions in designing your CBA – you should consult the literature for studies similar to yours and replicate their methodology where appropriate and possible.

Multicriteria Analysis

Cost benefit analysis still remains much to be wanted in an outcomes evaluation. Regardless of the health sector you are working in, it is likely that monetary evaluation methods won't satisfy your stakeholders. They will want to consider features of a program that are nearly impossible to quantify accurately. This is why multicriteria analyses is such a useful framework for evaluating programs and policies – especially when there are competing program to be chosen from.

Multicriteria analysis provides you with a methodology for making judgements by focusing you on highly specifiable, but pre-established criteria. The key here is that the decision framework is made ahead of time – before the criteria are evaluated. One feature that is sometimes, but not always, used is the addition of scores or weights to each of the criteria. Weights are generated according to the priorities of stakeholders – perhaps by asking stakeholders to rank them from highest to lowest and then adding the ranks of each stakeholder together to get a criteria-specific weight (See Table 11). Alternatively, you could ask participants to divide 100 points between each of your criteria and use the average scores as the value of the weights. (See Table 12)

With either method, participant anonymity had the greatest rank and cost had the lowest rank – reflecting the evaluation of the stakeholders of the relative worth of these criteria. You could also use established weights and established criteria frameworks – perhaps those identified through a literature search. Alternatively, you could avoid using

weights altogether. They are completely optional depending on what you and your stakeholders are looking for.

Table 11. Example method for creating weights

Rank-based Weighting	Cost	Time	Ease of Use	Participant Anonymity	Staff Support
Sponsor	5 - Most Important	1 - Least Important	2	3	4
Stakeholder 1	1 - Least Important	3	4	5 - Most Important	2
Stakeholder 2	1 - Least Important	3	5 - Most Important	4	2
Stakeholder 3	1 - Least Important	3	4	5 - Most Important	2
Summed Ranks	8	10	15	17	10
Value Allotment Weighting	Cost	Time	Ease of Use	Participant Anonymity	Staff Support
Sponsor	30	10	15	20	25
Stakeholder 1	5	15	20	50	10
Stakeholder 2	5	25	35	25	10
Stakeholder 3	5	20	30	35	10
Average Scores	11.25	17.5	25	32.5	13.75

In any case, the criteria selected by you and your stakeholders need to be measured. This means you will want to develop indicators that will help you make judgements about things like cost, time, ease of use, and whatever other criteria you land on. Common criteria in the health sector often include cost, time, health impact, patient/staff safety, patient acceptability, training requirements, staff support, and so on. Ultimately, any criteria that you feel would help discriminate between correct decisions and incorrect decisions about the value of a program.

After you have collected indicators for each criterion, you can assess how your program compared to either an alternative program or to not having the program at all. If you used weights, you might conduct sensitivity analysis to see if adjustments to the weights would lead you to make a different conclusion about a program. If you did not use weights, you can assess the various options qualitatively and see if one or two of them best meets the established criteria. Then, you can present this information to the decision makers and allow them to make a call regarding how they'd like to proceed.

Learning Activities

1. Think of a scenario that you would use for each of the following:
 - a. Multicriteria analysis
 - b. Cost-effectiveness analysis
 - c. Cost-benefit analysis
2. Establish a set of criteria that you think would be helpful in evaluating a program.
3. If you had to conduct a cost-benefit analysis. How would you monetize the value of a life? What considerations would you take into account? What challenges do you see in using this method of evaluation?
4. Under what conditions do you think it would be justifiable to operate a health program that is not cost effective or where the costs outweigh the benefits? Justify your answer.
5. The principal difference between cost-benefit analysis and cost-effectiveness analysis, is that CBA attempts to monetize (or quantify) the value of an outcome. Do you prefer this aspect of CBA compared to CEA? Explain why or why not.
6. Do you think the analytic methods discussed in this section improve how policies and programs are evaluated compared to those which only measure traditional health outcomes and impacts? If so, how?

Additional Readings

- Udvarhelyi et al. "Cost-Effectiveness and Cost-Benefit Analyses in the Medical Literature."
- Baultussen & Niessen. "Priority setting of health interventions: the need for multi-criteria decision analysis."
- March, Dolan, & Logon. "Prioritizing investments in public health: a multi-criteria decision analysis."

Part 3

How do I know my indicators are good?

Source: <https://www.spot-to-lab.fr/en/from-blood-sampling-to-dried-blood-spot-dbs/>

Measurement Theory

After selecting the criteria, you want to measure, the next big challenge is to be able to select indicators that measure the criteria of interest. In other words, it's time for measurement theory. Measurement theory is a theory that posits individual-level characteristics about an individual can be categorized and represented as data – often numerically. Generally speaking, you will have developed some sort of working model or theory that is guiding the work you do. You will use prediction to develop constructs that you think are part of your theory and then you will need to operationalize your constructs by creating items and indicators that relate to them (See Figure 12).

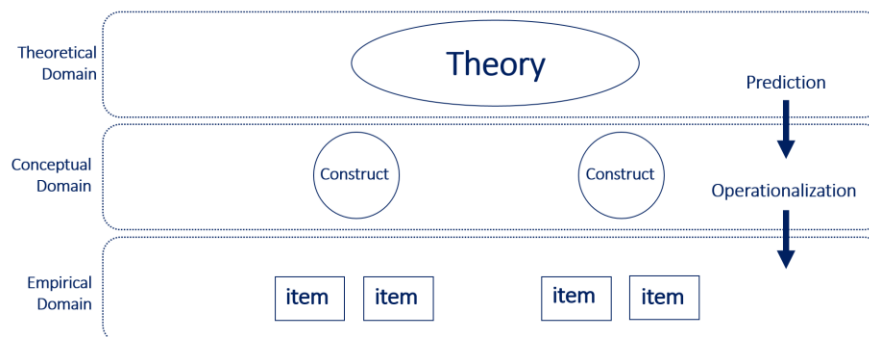


Figure 12. Measurement Theory

To conceptualize measurement theory, take a look at the image below. On the left you can see an environment with individual objects in it. One object is a green square oriented upward, one is a red circle oriented sideways, one is red circle oriented upwards, one is a blue triangle oriented upwards, and one is a blue circle oriented sideways. When measuring elements about this environment and the individual objects that occupy it, you have to make decisions on what to measure. Do you measure the axis tilt? The color? The shape? The location? You probably can only measure a few of these attributes. While these measurements can help you create a model – the model is imperfect. All models are imperfect. The goal of measurement is to develop models that are good enough to approximate what you are interested in.

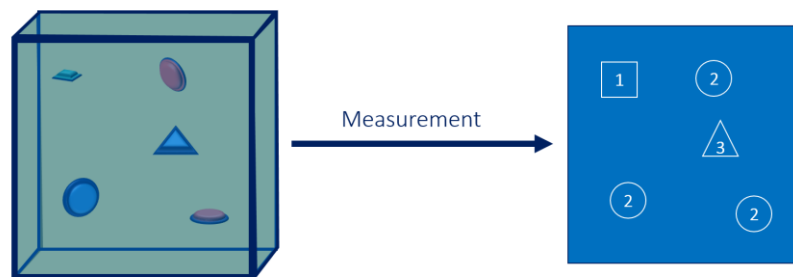


Figure 15. Measurement Example

After something is measured, you must also make inferences about your measurements. It should be your goal to make inferences that are valid reflections of what you observed. Statistical theory provides the basis by which researchers can make inferences (i.e., conclusions reached on the basis of evidence) from data collected for a specific purpose. To determine whether an inference is valid, you have to assess how good it is at measuring what you think it measures.

Validity

One measure of *indicator goodness* is validity. Validity has been described as the agreement between a test score or measure and the quality it is believed to measure. In other words, it measures the gap between what a test actually measures and what it is intended to measure. It is important that the measures you use are *valid* measures. Otherwise, your outcome evaluation will not be useful in helping you evaluate your

program. Validity can be measured in a variety of ways (See **Figure 15**), which we will discuss in this section.

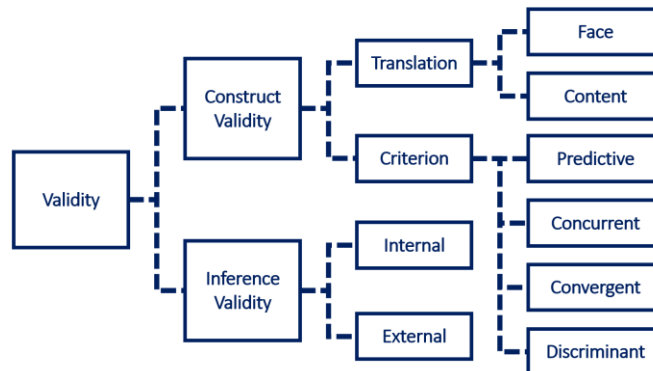


Figure 16. Types of Validity

Construct Validity

The first type of validity usually discussed is construct validity. Construct validity refers to the appropriateness of a variable being measured. Validity for constructs can be defined based on one's subjective evaluation of whether a measure matches the construct it is meant to measure (i.e. Translational validity) or how well the measure relates to other measures and characteristics (i.e. Criterion validity).

Translational validity is typically understood to include two sub-types and is often evaluated by so-called "experts." However, you can involve participants in this process using participant interviews, focus groups, and even survey questions. Doing so will help you to assess the two types of translational validity which you should be interested in when planning an evaluation:

- **Face validity** assesses validity on "face value." It asks whether the measure describes the construct well. You might assess the face validity of an indicator by asking somebody what they think it measures and whether they think there is a better way to measure what you are trying to measure. That said, a test can appear to be invalid but could still be useful in measuring your outcome of interest. This may be due to strong correlation between the construct being measured and the items used to measure it. However, a test that does not have face validity may confuse participants and others in your field may not be willing to use a test if it does not have face validity.

- **Content validity** asks whether a measure represents all the facets of a given construct. It might be assessed by asking stakeholders “What additional questions do you think are needed to measure this phenomena?” or “Is there any element of this phenomena missing from the questions I’ve provided?” In practice, few measures capture every dimension of a construct. Researchers therefore often rely on a limited subset of variables that capture each dimension “well enough.” That said, content under-representation occurs when important areas are missed and construct-irrelevant variation occurs when irrelevant factors contaminate the test. For example, if you want to know if someone is depressed, you might ask “are you depressed?” However, for content validity, you might need to ask several questions that get at the various parts of depression such as “Do you often feel sad?” or “Have you lost interest in activities you one enjoyed?”

Criterion validity is the second type of construct validity. It is often measured through empirical means – by comparing a measure or indicator to a “gold standard.” Criterion validity can be broken down into four subtypes:

- **Concurrent validity** is assessed by comparing two related measures completed at the same time.
- **Predictive validity** is assessed by examining the ability of a test to predict some event that occurs in the real world.
- **Convergent validity** occurs where measures of constructs that are expected to relate do so.
- **Discriminant validity** occurs where constructs that are expected not to relate do not, such that it is possible to discriminate between these constructs.

Inference Validity

Contrasted with construct validity, inference validity refers to the validity of a research design as a whole. In other words, it refers to whether you can trust the conclusions of a study. Generally, the issue under examination when assessing inference validity concerns causality. In doing so, it is important to recognize that statistical measures show relationships, but it is the theory and the study design that affect what kinds of claims to causality you can reasonably make based on the results of any given study.

There are two types of inference validity that we are generally concerned with: Internal validity and External validity.

- **Internal Validity** refers to whether conclusions, especially relating to causality, are consistent with research results (e.g., statistical results) and research design (e.g., presence of appropriate control variables, use of appropriate methodology). In other words, it asks whether your interpretations are consistent with your results and whether you have given sufficient weight to the limitations of your study. To strengthen internal validity, you will want to carefully think through the potential sources of bias that might be affecting your study; ask yourself if you have controlled for all the relevant and important confounders; and make sure that you have conducted the right statistical tests and interpreted those tests correctly.
- **External Validity** Refers to whether the results of a study can be applied, or *generalized*, to the real world. Three strategies for strengthening external validity:
 - **Sampling.** Select cases from a known population via a probability sample, then claim the results apply to the population as a whole.
 - **Representativeness.** Show the similarities between the cases you studied with a population you wish your results to be applied to.
 - **Replication.** Repeat the study in multiple settings. Use meta statistics to evaluate the results across studies. Although journal reviewers might not agree, consistent results across many settings with small samples may be just as good (or better) than a large sample of a single settings.

Reliability

In addition to validity, indicator goodness can also be evaluated based on the reliability of a measure. There are two types of reliability that we are generally concerned with: Stability and Consistency.

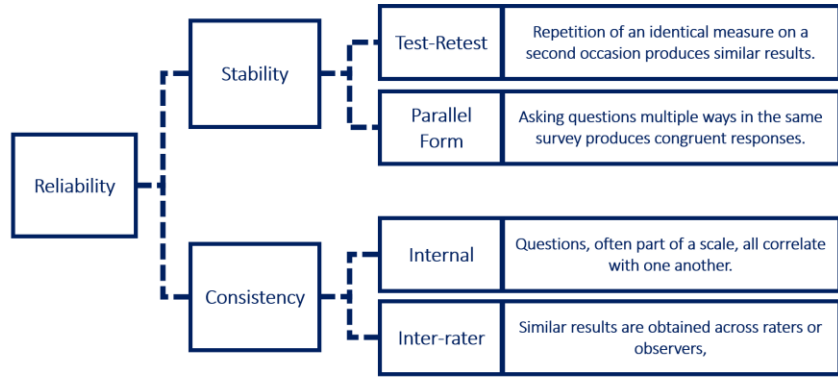


Figure 17. Types of Reliability

Stability

Stability is a measure of the repeatability of a test over time, that it gives the same results whenever it is used (within defined constraints, of course). Two types of stability are usually measured: Test-Retest reliability and Parallel Form reliability. We estimate test-retest reliability when we administer the same test to the same sample on two different occasions. This approach assumes that there is no substantial change in the construct being measured between the two occasions. Variation should be due to the test, not to any other factor. That said, it is generally accepted that over long periods of time, stability decreases due to factors other than those a scale or instrument measures. Several factors may lead to poor measures of stability. Among these, carry-over effects (i.e., people remember answers from last time), practice effects (i.e., repeat test taking improves scores), and attrition effects (i.e., some participants don't return for re-test) all have the potential to bias test-retest comparisons.



Figure 18. Test-Retest Reliability Diagram

Parallel form stability, also referred to as equivalence, differs from test-retest reliability in that two measures are administered at the same time and the results of each are compared. Often times the two versions of the scale are created from a larger set of questions that address the same construct. Items from the larger set are randomly decided and both are administered to the same people. The correlation between the

two forms estimates the stability of the measured construct. Similar approaches are used when you want to shorten a scale.

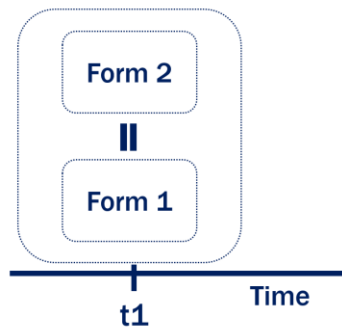


Figure 19. Parallel Form Reliability Diagram

Consistency

Consistency is a measure of reliability through similarity within the test, with individual questions giving predictable answers every time. The first type of consistency is inter-rater reliability. There are two major ways to actually estimate inter-rater reliability. If your measurement consists of categories – the raters are checking off which category each observation falls in – you can calculate the percent of agreement between the raters. The other major way to estimate inter-rater reliability is appropriate when the measure is a continuous one. There, all you need to do is calculate the correlation between the ratings of the two observers.

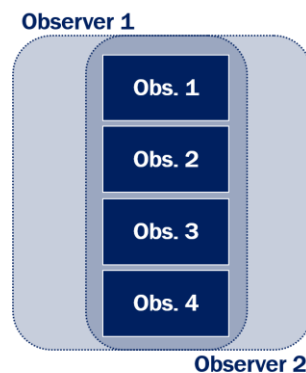


Figure 20. Inter-rater consistency Diagram

Another form of consistency is internal consistency. Internal consistency estimates may be affected by the difficulty of the test, the spread in scores, and the length of the examination – but in general items within a scale should lead to similar conclusions

about how an individual scores on an overall construct. There are many measures for internal consistency, including the following:

- **Average inter-item correlation** tells you the average of all correlations.
- **Average item-total correlation** tells you the correlation between each item and the overall scale score.
- **Split-half reliability** is calculated by randomly dividing all items that purport to measure the same construct into two sets. You then administer the entire instrument to a sample of people and calculate the total score for each randomly divided half. The split-half reliability estimate is simply the correlation between these two total scores. The split-half approach is very similar to the parallel reliability described earlier. The major difference is that parallel forms are constructed so that the two forms can be used independent of each other and considered equivalent measures. A parallel form may sometimes include an earlier version or a longer version (and, as such, is very similar to convergent validity). A problem with this is that the resultant tests are shorter and can hence lose reliability. Split-half is thus easier to use with longer tests.
- **Cronbach's α** is a lower bound estimate for the reliability of a scale calculated as a function of the number of items in a test, the average covariance between item pairs, and the variance of the total score. It is widely used and reported. When interpreting Cronbach's α , alpha increases as the intercorrelations between items increase. It is mathematically equivalent to the average of all possible split-half estimates. Chronbach's α values ≥ 0.9 are excellent, ≥ 0.8 are good, ≥ 0.7 are acceptable, ≥ 0.6 are questionable, and < 0.6 are poor.
- **Kuder-Richardson** is another measure of internal consistency reliability for measures with dichotomous choices. It is a special case of Cronbach's α in which high values (between 0 and 1) indicate that the examination is more likely to correlate with alternate forms (a desirable characteristic).
- **Intraclass correlation coefficient (ICC)** are widely used reliability indices in test-retest, intrarater, and interrater reliability analyses. There are 10 forms of ICCs. Because each form involves distinct assumptions in their calculation and will lead to different interpretations you should investigate these carefully before use. However, when interpreting these results, the lower

bound of the 95% confident interval of the ICC estimate indicates excellent (0.90), good (0.75), moderate (0.50) and poor (0.50) reliability.

Now that you have a bit better grasp on reliability and validity you should get a sense of what these measures mean for the indicators and instruments you select for your outcome evaluation. **Table 21** shows a helpful two-by-two table on how to think about your measures once you have assessed their reliability and validity.

Table 21. Relationship between Reliability and Validity

	Valid	Not Valid
Reliable	You are measuring what you think you are using a measure that will produce stable and consistent results.	You have a reliable measure of something, just not what you think it is.
Not Reliable	The average measurement is right on, but each individual measurement has error and is un-usable by itself.	If you are measuring something, its not what you want and its not a reliable way of measuring whatever it is measuring.

Considering Alternative Causal Explanations

In addition to challenges with validity and reliability, even good measures can be confounded by things outside of your control. For instance, if there is too much variation in your data, or your subjects are poorly selected, or there are complex interactions across constructs – these and many other factors have the potential to bias your results. Statistical significance alone, even when supported by measures of reliability and validity, should not be taken at face value. Indeed, you must also think about threats to internal validity, including those mentioned below:

- **Observer** effects occur when participants behave differently because they know their behavior is being measured.
- **History** can impact your validity when other influential factors occur during the same time of an intervention (e.g., food shortage, job loss). Using a control group can help address this challenge.
- **Maturation** arises from automatic improvements that accompany the passage of time. For example, a child's height or grieving. Address this threat by using a control group and a pre/post test.
- **Testing** effects appear after repeated test taking individuals respond carelessly or as they think they are expected to. Conducting post tests (when you randomized your subjects into the intervention) can help with this.

- **Instrumentation** challenges arise when test administration is not consistent. To address this threat, simply trained survey administrators to make sure there is consistency in the ways questions asked.
- **Selection bias** can impact validity if the outcome is related to who you allow to be part of your study. You can address this through carefully thinking about your inclusion criteria and by randomizing included participants into the control and intervention groups of your study.
- **Loss to follow-up** occurs when participants loose contact with the study. To address this, use large sample sizes, follow-ups, and incentives.
- **Placebos** play a role when positive effects emerge simply due to exposure to an intervention and not to the intervention itself. Control groups (either placebo, attention, or gold standard) can minimize this threat.
- **Contamination** can occur when the comparison group is aware of the intervention and changes their behavior for better (decreasing the effect size) or worse (increasing the effect size).
- **Regression to the mean** occurs when you sample participants with high or low scores within the general population. Scores in these subjects tend to regress back towards the mean as their extreme values are often anomalous.

I know that it can get a bit overwhelming to think about how you can overcome all of these threats to validity. Do not think that every evaluation needs to address every issue. It is true that higher quality evaluations will follow some of the gold standard recommendations that we will discuss in the following chapters, but it is not always possible to design a perfect evaluation. You simply have to try your best. When you are unable to address a bias, however, do not try to hide it. Speak about it openly – make sure your stakeholders are aware of the bias. It is not good for them to be surprised at the end of an evaluation either – so make sure you try and address these concerns with your stakeholders well before the results of your evaluation are made available.

Often budgets are one of the limiting factors in your ability to control for these threats to validity. Evaluations regularly receive too little funding to fully address the scope of questions they are asked to. Yet, the focus on “implementation science” in Canada is opening up new opportunities to conduct evaluations of this sort. So always be on the lookout for grants that specifically aim to understand the process of implementation – particularly when there is already an evidence base established for an intervention.

Learning Activities

1. Search out the diagnostic criteria for depression from either:
 - a. The Diagnostic and Statistical Manual of Mental Disorders, 5th Addition (DSM-IV)
 - b. The International Classification of Disease, 10th Revision (ICD-10)After reviewing these criteria, take a look at one of the following scales and discuss what changes you might make to improve its validity or reliability.
 - i. Hospital Anxiety and Depression Scale (HADS)
 - ii. Center for Epidemiologic Studies Depression Scale (CES-D)
 - iii. Beck Depression Inventory (BDI)
 - iv. The Major Depression Inventory
 - v. Hamilton Depression Rating Scale (HDRS)
 - vi. Montgomery-Åsberg Depression Rating Scale
 - vii. Patient Health Questionnaire (PHQ)
 - viii. Primary Care Evaluation of Mental Disorders (PRIME-MD)
2. Take any two of the scales above and see how closely your scores align (divide your score by the total score). Use any standard cut offs and see if the scales classify you the same way or not.
3. Identify the potential validity threat and how you would address it in each of the following scenarios:
 - a. You find out that the government introduces an anti-poverty program at the same time you open a test kitchen that you are about to evaluate.
 - b. You feel that it is impossible to keep those in the treatment group separate from the control group.
 - c. Your data will be collected by many individuals with varying level of experience with data collection.

Additional Readings

- Fitzgerald et al. ["The Reliability and Validity of a Brief Diabetes Knowledge Test."](#)
- Krieger et al. ["Experiences of Discrimination: Validity and reliability of self-report measures for population health research on racism and health."](#)
- Bunge et al. ["Reliability and validity of health status measurement by the TAPQOL."](#)
- Otieno-Odawa et al. ["Validity and reliability of data collected by community health workers in rural and peri-urban contexts in Kenya."](#)

Chapter 5

QUANTITATIVE DATA COLLECTION

Learning Objectives

- Design a quantitative study.
- Develop a questionnaire.

Part 1

How do you design a quantitative evaluation?



Source: <https://www.analyticsindiamag.com/5-key-personality-traits-every-analytics-leader-should-have/>

Quantitative Evaluation

This chapter, and the four that follow it, focus on how to collect and analyze data from the participants you recruit into your study. There are two broad categories of studies that are widely used in health research: quantitative studies and qualitative studies. Often times you will hear people describe themselves as “qualitative” or “quantitative” researchers. However, this dichotomy should be more about the types of research questions individual researchers than it is about identity as a “numbers” or “word” person. Indeed, as an evaluator you will need to employ both research methods depending on what your goals and objectives are.

Use of Quantitative Methods

Quantitative research methods tend to be useful when you are interested in the perspectives of individuals, when you are hoping to get a large number of individuals to provide information to you, and when you are fairly certain what constructs you would like to measure. As we will talk about in the chapter on mixed methods, quantitative methods can be used hand in hand with qualitative methods as a way to iteratively develop your understanding of a particular aspect of the project you are working on.

When conducting a quantitative research study, there are a variety of decisions for you to make. Largely these decisions will be impacted by how many resources you have to throw at your evaluation and how interested you are in controlling for the threats to validity discussed in the last chapter. You likely learned about study designs – and many of their key features – in an introductory epidemiology or biostatistics course. Epidemiology often focuses on cross-sectional, case-control, and cohort studies. However, in health services research this trifecta approach is replaced with a more detailed and applied set of study designs that are – to varying degrees – suited to study the development of a program and draw comparisons between intervention and control groups.

One of the most common study designs is a **one-group post-test only** design. It is popular because it is a lower cost and less resource intensive study design. In this design you do not have a control group and you do not have a pre-test. However, this is not to say that you are not able to use this study design to make comparisons. You might use it to understand for whom your program is working for by looking for characteristics of participants who did not perform as anticipated. Given that you do not have a pre-test to compare it to, the outcomes of these one-group post-test only designs are often dichotomous or focus on achieving certain outcomes. For example, you would not be able to look at scores at the end of an intervention working to reduce loneliness using some sort of loneliness score – because you don't know how individuals would have scored before hand. However, you might be able to understand whether or not somebody found your intervention acceptable or use it to identify parts of your program that need to be reconsidered.

In a variety of scenarios, it is useful and fairly cost-effective to add a pre-test, making your study a **one-group pretest post-test design**. The advantage with this method is you can compare individuals at the end of the study to their status at the beginning of your study. For instance, if you were evaluating the effectiveness of methadone to reduce cravings for heroin among heroin users, you could use a scale that measures cravings at the beginning and end of the study and compare individuals to themselves at each time point. While you can't necessarily prove that these individuals improved because of your intervention (because you lack a control group), you can at least assess whether or not individuals even improved. Thus, the one-group pretest-post-test design is useful for a wider range of outcomes – particularly those that are not clear-cut statuses.

Similar to the one-group pretest post test design, **interrupted time series designs** involve multiple measurements, usually several before and several after the intervention. The advantage of this design is that even though you may not have a comparison group, you can look to see whether or not the change associated with the intervention is attributable to changes that are already taking place before and after the intervention. This is a great method for evaluating the impact of a policy, where all those in a population might be affected and it is therefore difficult to organize a control group. Individuals, in a way, can act as their own controls.

Advancing beyond the one-group study design is the **post-test only with non-equivalent groups design**. This design lacks randomization, but has a control group. As such, you can make comparisons between participants of a study and non-participants but it is not entirely clear whether the differences between the groups is solely attributable to your intervention – because the groups in the first place are non-equivalent. This study design is often useful when you want to have a control group, but you are not able to randomize participants into the program – perhaps for ethical reasons. For example, if you were evaluating a smoking cessation program you probably would probably prefer to let participants choose to be in the program or the control arm of the study. However, differences in people who participate in the study compared to those who do not are difficult to quantify and may confound the conclusions of your study. Imagine, for instance, if people who chose not to participate in your study were more likely to feel they could just quit on their own and that these people’s increased self-efficacy actually made them more capable of quitting. Your study may reveal equal or higher cessation in this group – leading you to believe that your program is not effective when in fact self-efficacy is a confounding factor in this conclusion.

Taking this design one-step further, a **non-equivalent control group design** adds a pre-test to each of the groups. This can allow you to see if the improvements in the control group differ from those seen in the intervention group. The design is still not randomized, but it does allow you to definitively say whether improvements in the groups were made and whether they actually differed.

If you add randomization to a non-equivalent control group design, you have created a classic **experimental design** – also referred to as a randomized control trial. In this study groups are made equivalent by randomizing potential confounders between the two groups. However, this is not always possible, it tends to be more expensive, and it tends to be a bit more time consuming. Nevertheless, you can be *fairly confident* that

differences in improvements between the groups is likely due to the intervention and not to some other factor.

However, what if you have a randomized group, and you just eliminate the pretest? This is called a **post-test only control group design**. You would want this if the criteria for inclusion is standard across your groups and participants can be randomized into the intervention or control arm of the study. In this case, the pre-test status is irrelevant and you are only interested in the post-test. You might also use this design if you are concerned that a post-test could negatively impact your post-test results. For example, if you ask participants if they are aware of a new drug, they might go and learn about the new drug regardless if they are in the intervention or control arm. Therefore, a question assessing the study's ability to increase your awareness of the new drug would be invalidated by your pre-test.

In addition to the pre-test post-test control group design (i.e., the classic experimental design), a **cross-over study** is another study design that carries the weight of an experimental design. In some ways it is similar to the time series design in which a series of measurements are made over time. However, instead of randomizing participants to a control group or intervention group, you would randomize the sequencing of treatments and/or placebos they received and they would act as their own controls. The benefit of this approach is that you are able to control for confounding because most characteristics stay the same within a person except those that are altered by the intervention.

Finally, one of the most robust (but expensive and time consuming) research designs is the **Solomon four group design**. The Solomon four group design involves two groups who receive the intervention and two groups who do not. Two of the groups receive a pre-test, while the other two do not. This allows you to identify the effect of repeat testing – which as discussed earlier can impact your results if participants are changing as a result of the pre-test.

Now that you have been introduced to these designs, the big take home message should be the utility of using (1) control groups, (2) randomization, and (3) and sequencing as tools to guide your design and control threats to validity.

Learning Activities

1. Explain how the following aspects can be used, individually and together, to control threats to validity.
 - a. Control groups
 - b. Randomization
 - c. Sequencing
2. If classic experimental studies are the best way of determining causality, why would we want to use the other kinds of studies?
3. A methadone clinic reports to you that they have found that participants who drink alcohol are more likely to overdose in the clinic. They want to know whether consuming alcohol leads to greater overdose risk for methadone users. Describe a study design that might be used to investigate this issue.
4. An evaluator tests all of their subjects, gives them a treatment, and then tests them again. What type of quasi-experimental design is this an example of? What threats to validity might impact this study design?

Additional Readings

- Campbell and Stanley. (1959). "Experimental and Quasi-Experimental Designs for Research."

Part 2

How do I develop questions?

Source: <https://blogs.ei.columbia.edu/2017/09/21/quiz-how-much-do-you-really-know-about-climate-change/>

Question Development

Once you have figured out your study design, the next step is to begin developing the indicators and questions that you will use in your data. As a first step you will need to develop a research question. A research question is the goal of a research project and aims to answer a **specific question** about your **topic of interest**.

To create a research question you should follow the steps diagrammed in Figure 21.



Figure 21. Steps to developing a research question.

Taking this approach will help you develop clear research questions. Clear research questions, in turn, also make it easier for you to collect and analyze data. Take for

instance the example in Figure 22, which shows the narrowing of a research question from a general research topic.

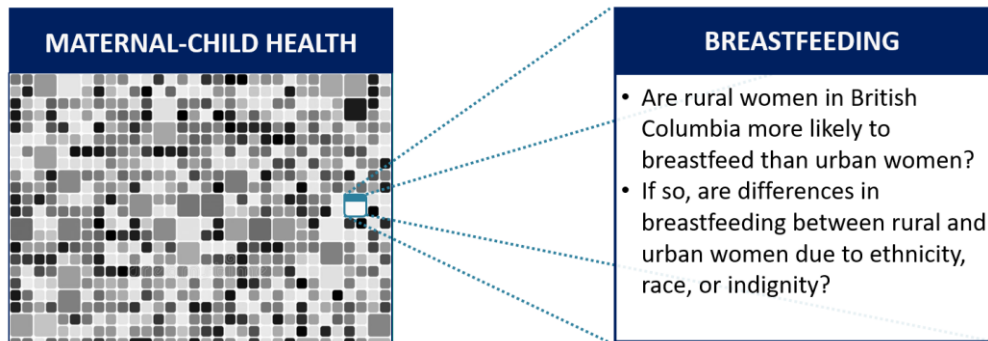


Figure 22. Developing a research Question.

Often times in an evaluation study, the development of a research question is rather easy – largely because the stakeholders already know they want or need to know and will tell you what they are looking for when they engage with you in preparing for the evaluation. So once you have your research question developed, you have to begin thinking about the type of data you will want to use to answer your question.

Data Types

You are probably familiar with the various data types available to you (See figure 23).

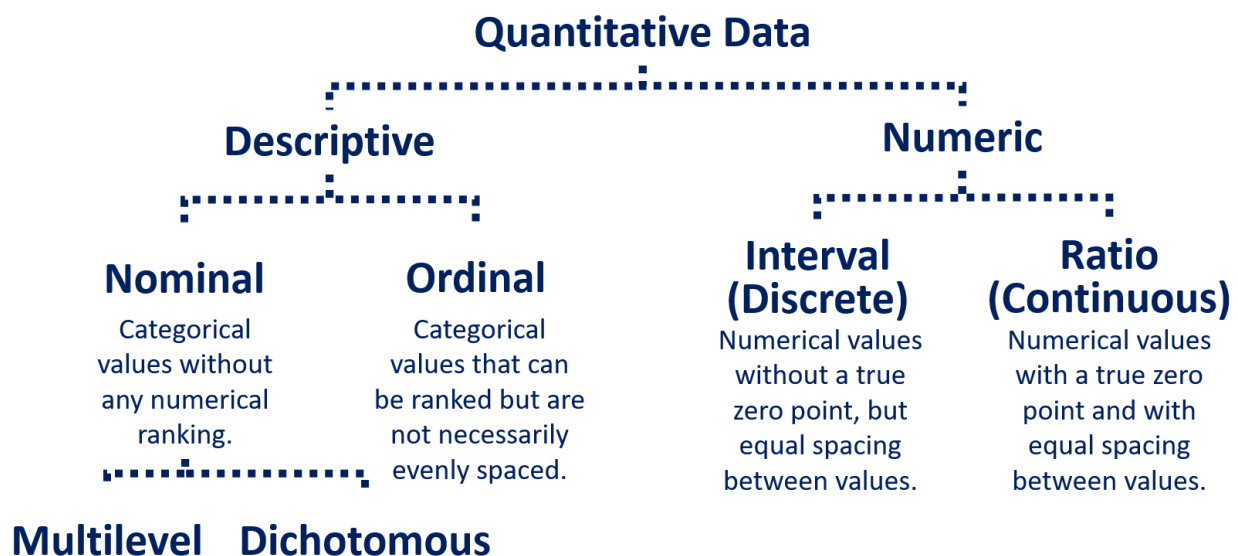


Figure 23. Data Types

One of the key things to remember about these data types is that each one preserves some quality of the “real” thing you are trying to measure. These qualities are described as properties – and there are generally thought to be 5 properties that can be captured in qualitative data:

- **Equality** properties allow you to compare equalities between objects. For example, if object A is 560° Kelvin (K) and object B is 280°K then value A will not equal value B.
- **Ordinality** properties allow you to compare the order or ranking of objects. For example if object A is 560°K and object B is 280°K then value A will be greater than value B.
- **Interval** properties allow you to compare the intervals between objects. For example, if object A is 560°K and object B is 280°K and object C is 140°K then the interval between value A and value B will be twice as great as the interval between value B and value C (i.e., each 1-unit value increase will equal a 140° Kelvin increase).
- **Ratio** properties maintain the correspondence between the ratios of the measured values to the ratios of the actual properties being measured. For example, if object A is twice as hot as object B then value A will be twice as high as value B.

Ratio scales maintain equality, ordinality, interval, and ratio properties; Interval scales maintain equality, ordinality, and interval properties; Ordinal scales maintain equality and ordinality properties; and Nominal scales maintain only equality properties.

Once you have figured out what type of data you want to use, you will be able to finally start developing survey questions to measure these outcomes. However, developing survey questions is not as easy as simply throwing some words together. You must account for the fact that human beings will be the ones completing your survey. As such, good survey questions start with good spelling, grammar, and punctuation. They should be readable to someone with approximately a 6th grade reading level, avoid words with more than two syllables, avoid jargon, and when jargon is not avoidable they should provide explanations for terms that are not universally well-known. Questions should be short, simple, and direct.

If questions are not easily understood, participants may drop out of the survey, try to guess what you are asking and answer *that* question, or select random answers. In other words, good questions get you good data, and bad questions get you bad data. Bad questions come from questions that require unnecessary cognitive effort.

Question Types

In quantitative surveys, there are five main question formats used.

- **Fixed choice questions** include questions such as “Are you happy with the care you received from your doctor” with the response options fixed as either “yes” or “no.” These fixed choice formats enable easier data collection and analysis, and are a good approach for large populations. While they provide a variety of ways to ask questions, you must include all reasonable possibilities or you may not get a true insight into their response. While it is tempting to come up with some categories and then to simply allow participants to choose “Other” and write in a response, the written-in responses are rarely usable because they lack sensitivity or specificity (i.e., participants who simply didn’t think to write in an other response might qualify for the other category if they had thought of it). Usually, the only use of the “other” category is to identify situations in which participants fit into one of the original fixed response options. Sometimes response options in a fixed question are randomized to reduce bias that might emerge from participants clicking through the survey quickly without reading the questions carefully. You can also include, “refuse to answer” or “I don’t know” if you think a participant might otherwise leave the response blank. This is important in paper and online surveys when response to questions is not forced because if a question is simply left blank it is not clear whether it is blank because a participant didn’t see the question or because they did not feel they fit into the category.
- **Open ended questions** are sometimes used in surveys. While these are qualitative question types, I mention them here because I want to make it clear that surveys can be used to collect qualitative data. Including these questions offers flexibility and freedom to those completing your survey. As such it can result in rich and interesting data. That said, these questions tend to be a bit more time consuming, can be difficult to analyze if you have many respondents, and can result in vague or confusing responses.

- **Check-all-that-apply questions** are similar to fixed response option questions, except an individual is able to choose more than one option. For example, I might ask participants “Which of the following services are you interested in accessing?” and then provide a long list of services. This makes it easier for participants to respond to questions. However, as with other blank options, it is difficult to say with a check-all-that-apply question whether a check box is left unclicked due to it not applying to the participant or because the person skipped over the question or didn’t see a specific response option.
- **Matrix style questions**, like check-all-that-apply, allow you to measure responses more easily for multiple questions or themes. An example of a matrix question can be seen in **Figure 24**. The benefits of this question is that it can save the respondent time and space on a survey if multiple questions use the same scale or response options. On the other hand, they can be overwhelming – especially if you have an entire “wall” of questions in this style or when there are lots of points on the scale. I recommend that generally speaking using four to seven response options is best, with each point clearly defined. Also when using Likert-scales in this format, be sure that the question aligns with the response option choices. For example, if you are asking about disagreement, the response options should be “strongly agree” to “strongly disagree.” **Figure 25** shows several multi-point scales used in health research.

Example

Below is a list of ways you might have felt.
How often have you felt this way during the past week?

	Rarely	Sometimes	Often
Bothered by things	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Loss of Appetite	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>
Hopeless	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fearful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Restless	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Lonely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 24. Matrix Question

Agreement	Importance	Quality	Satisfaction	Likelihood	Frequency	Frequency
Strongly Agree	Very Important	Very Good	Very Dissatisfied	Definitively	Always	Very Frequently
Agree	Moderately Important	Good	Moderately Dissatisfied	Very Likely	Very Frequently	Frequently
Slightly Agree	Somewhat Important	Acceptable	Slightly Dissatisfied	Somewhat Likely	Frequently	Occasionally
Slightly Disagree	Somewhat Unimportant	Poor	Slightly Satisfied	Equally Likely	Occasionally	Rarely
Disagree	Moderately Unimportant	Very Poor	Moderately Satisfied	Somewhat Unlikely	Rarely	Very Rarely
Strongly Disagree	Very Unimportant		Very Satisfied	Probably Unlikely	Very Rarely	Never
				Definitely Not	Never	

Figure 25. Likers Scale Response Options

Pitfalls to Avoid

When developing questions, it is good not to use too many types of questions. Remember you do not want it to be laborious for participants to complete your survey. Additionally, there are many “pitfalls” that we tend to recommend against when developing questions. We will describe these pitfalls in this section:

- **Avoid unnecessary questions.** If one question captures the same information as the other, keep whichever is more informative. Do not ask questions that are merely interesting – ask questions that are directly related to your research.

Example

Do you use condoms when having sex?

Yes

No

How frequently do you use condoms during sex?

Always Most of the time Some of the time

Seldom Rarely Never

- **Avoid the bells and whistles** (e.g. sliders, heat maps, animations) unless they are absolutely necessary. Sticking with the question types discussed in the last section will generally be easier for your respondents and will work on more devices.
- **Double Barrelled questions** are questions that ask about multiple things in a single question. Sometimes a respondent’s answer might differ for each sub-component of the question. Generally speaking, double- barreled questions can be identified by the presence of and/or clauses.

Example

How satisfied are you with your doctor and nurse?

Very Satisfied

Somewhat Satisfied

Somewhat Dissatisfied

Very Dissatisfied

- **Inadequate response options** do not include all possible/appropriate responses.
- **Open ended “other” options** should be used sparingly, but as necessary. (Sometimes use even if data wont be).
- Avoid use of **double negative** and **negatively framed** questions as these can be difficult for readers to understand.

Example

Are you **not** happy with the care you received from your doctor?

Yes
 No

Are you happy with the care you received from your doctor?

Yes
 No

- **Leading Questions** prompt or encourage one answer over another through subtle or direct coercion.

Example

Do you believe that **Truvada is an effective way to prevent HIV?**

Yes
 No

How would you rate the effectiveness of Truvada in preventing HIV?

It is effective
 It is not effective.

- **Loaded questions** include complex, unjustified assumptions that may not be interpreted universally.

Example

In the past six months have you had **unsafe sex?**

Yes
 No

In the past six months have you had sex without a condom?

Yes
 No

- **Absolutist Questions** require participants to make absolutist claims.

Example

Do you **always** tell the truth?

Yes
 No

Do you agree with the following statement?
 “It is **never** okay to lie.” Yes No

- **Assuming Questions** require special knowledge or understanding, these should be described to the reader, if possible.

Example

Does Bill C-130 benefit your family?

Yes

No

- **Insensitive Questions** may not be culturally appropriate to the audience answering them.

Example

What is your gender?

Male

Female

- **Non-Exclusive responses** “overlap” making it difficult for individuals to select 1 answer.

Example

What is your relationship status?

Married

Divorced

Single

Widowed

What if I am widowed, but re-married?

Divorced and now single?

- **Middle response options** can act as “cop out” or “does not apply to me” choices. Avoid using a neutral option in a scale. These are difficult to classify (e.g., is it agreement or disagreement). Further, individuals who might otherwise make a choice, may choose a middle option instead. In these cases, it might be better to make participants “choose a side.”

Example

How likely do you feel it is that you will get HIV?

Strongly Agree

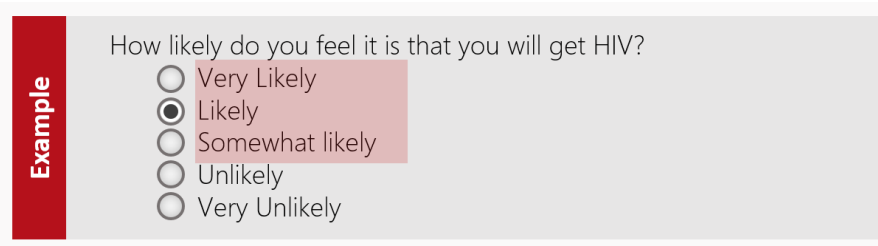
Agree

Neither Agree Nor Disagree

Disagree

Strongly Disagree

- **Unbalanced scales** are scales with an unequal number of options for both sides of the spectrum.



The image shows a survey question: "How likely do you feel it is that you will get HIV?". Below the question are five radio button options: "Very Likely", "Likely", "Somewhat likely", "Unlikely", and "Very Unlikely". The "Likely" option is selected, indicated by a filled radio button. The options "Very Likely" and "Somewhat likely" are highlighted with a light red background, while the other three options are on a plain white background. A red vertical bar on the left side of the question area contains the word "Example" in white text.

Avoiding these pitfalls will make it easier for your participants to complete your survey and will reduce the cognitive effort required for them to provide the data you need for your study.

Survey Presentation

However, questions alone do not impact how easy it is for your participants to complete a survey. It is also important that the survey is well presented. One of the most important factors in presentation is length. Avoid long surveys. Focus groups suggest participants do not want to take more than 10 or 15 minutes doing a survey. While you can't achieve this goal for some research questions, you should delete duplicate or vague questions and employ *skip logic* or *display logic* so respondents only see the questions relevant to them. Further, by adding white space and decluttering your survey, you can give your survey a clean look that reduces the cognitive load. If you are presenting a survey online, use a progress bar so participants know how far along they are. Break up long pages into bite sized chunks to eliminate how much participants have to scroll. Don't force participants to track their eyes across the entire screen, increase the margin size if you are using an online survey or split the questions into columns if you are using a paper survey.

As mentioned above, skip logic and display logic are done important tool for improving the flow and presentation of your survey. **Skip logic** Instructions (either on paper or programmed surveys) direct a respondent to a particular question based on their answer to a previous item. Skip logic involves the use of gateway questions to make sure participants aren't asked irrelevant questions and to prime participants to better recall and probe their experiences. When using skip logic, organize your questions from general to specific. (e.g., Do you have a primary care physician When was the last time you visited your primary care physician? What was the purpose of your visit?"). Skip and

display logic not only improve survey aesthetics but they also help minimize data errors as well because people cannot answer questions that do not apply to them. In paper based surveys, even when skip logic is clearly instructed (i.e., “If yes, skip to question 4), participants will often complete questions that should have been skipped and it will be difficult to assess data quality when answers in different parts of the survey conflict with one another.

Example

4. “Are you not happy with the care you received from your doctor?”

Yes

No [Skip to Question 10]

Display logic is like skip logic, but is a feature unique to online surveys. Display logic involves programming your survey so that questions are displayed to respondents only if they meet a set of predetermined criteria. Predetermined criteria may include:

- Responses to one or more earlier items (e.g., “If yes to Q1 and no to Q2 and yes to Q3: show Q4”)
- Data associated with your sample (e.g. All freshmen see Q1, all sophomores see Q2, all juniors see Q3.)
- Randomization (e.g., participants are randomized to see two versions of the same question).

Example

1. Were you wearing a shirt?

Yes

No

2. Were you wearing shoes?

Yes

No

3. Did you attempt to get service?

Yes

No

Q4 displayed if

- Q1 or Q2 = No; and
- Q3 = Yes

4. Were you denied service?

Yes

No

It is important to recognize that skip logic creates meaningful “missingness” in your data. For example, if you are surveying people who use drugs and people who do not, those who report not using any drugs in a gateway question will have missing responses

for any of the questions relevant to drug use. Often times, special codes such as 9999, 8888, 7777, 6666. Each code is assigned a special value that is reported in your data dictionary.

ID	SHIRTS	SHOES	ATTEMPT	DENIED
1	Yes	Yes	No	8888
2	No	Yes	Yes	No
3	No	No	Yes	No
4	Yes	No	No	8888
5	Yes	No	Yes	Yes
6	Yes	Yes	Yes	8888
7	No	No	No	7777

9999 = Missing

8888 = Question not asked

7777 = Participant Selected "Refused"

6666 = Participant Selected "Don't Know"

Closely linked too the issue of display and skip logic, is the use of randomization in the order in which questions are presented. Randomizing the order in which questions are presented can reduce bias – particularly if the act of asking some questions might influence how participants answer other questions. When employed along with skip logic, sometimes it is necessary to randomize “blocks” of questions or questions within blocks, which can be useful in allowing you to govern the overall flow of a questionnaire. When doing so, start with exciting and interesting material, not demographics. This will help participants engage with your questionnaire. This does not mean that demographics have to be saved until the end, just don’t make them the very first thing you show people. Indeed, if people are going to refer their friends to participate, make sure you finish with some interesting stuff as well. Likewise, if questions are really important (as demographics often are), they should appear earlier, to avoid drop off.

Most surveys also include a title, purpose statement, letter of implied consent, information about how their data will be kept confidential, and instructions for how to complete the survey and receive their honorarium (if applicable). You should also make sure that it is clear how to complete a question. Avoid asking participants to “circle” options, use checkboxes or radial dials instead. Questions are often bolder while response options are not – this makes it easy to flow from question to question. Your questions and response options across questions should be aligned for a clean look. When using scales, do not switch which direction the scales are going (e.g., put agree consistently on the left or right, opposites of disagree). It is good for related questions to be closer together, unless you feel the need to randomize the questions in order to

reduce bias. It is best to print paper surveys on only one side of the paper and to use pastel colored papers (as opposed too white). If you are using an online survey, you will have to decide whether participants can go back once they have answered a question and whether they can save their progress and return later. Communicating whether or not these options are possible is important to ensure participants don't loose their data because they misunderstood.

Once you have done everything you can to make a good questionnaire, the only thing left to do is to test your survey (and test, and test again, and test some more). Circulating your survey to friends and colleagues, and, when possible, the population you are planning to work with is a key step to preparing your survey. You can ask those piloting your survey whether the survey makes sense, whether people understand what you are asking, whether the response options make sense, whether the purpose of the survey is sufficiently transparent, and how they feel about the survey process. You can gauge whether they feel like their responses are contributing to something important and whether the survey length is appropriate given any incentives provided. They can help you improve the look of the survey, declutter areas that feel a bit overwhelming, and test the skip and display logic to ensure it is working correctly (online) or clear enough to follow (on paper).

Once your survey is finalized, you should be ready to translate it into any languages that are needed. In Canada, common languages may include English, French, Mandarin, Cantonese, and Punjabi – though you might choose the languages that are most appropriate for your study population. Often times you higher a translator who will conduct a forward translation. The goal of the forward translation is not to reproduce a word-for-word translation of your survey. Rather, the goal is to relay the conceptual equivalencies in each translation. After your survey has been translated into a target language, you should have an expert panel back translate the survey into its original language. Doing so will tell you whether the meaning of your questions was maintained throughout the translation process. After making any needed revisions, you can then pre-test and pilot your survey in the new language – just like you had done in the original version.

Learning Activities

1. Earlier we used temperature of objects as an example for the various qualities data have. In these examples, it should be recognized that Kelvin possesses ratio, interval, ordinal, and equality properties (i.e., 286 is 47 times larger than 6; 548 is 12 times larger than 44); Fahrenheit possesses interval, ordinal, and equality properties; and Celsius possesses interval, ordinal, and equality properties. Given this information, what kind of data types are Celsius, interval, and Fahrenheit?
2. Review this section and create a checklist of all the recommendations for designing surveys and survey questions. Select one of the example surveys below and use your checklist to evaluate the survey.
 - a. [Sex Now Survey](#)
 - b. [University of Barcelona English Learning Questionnaire](#)
 - c. [Finnish Questionnaire on the English Language](#)
 - d. [ESPAD Substance Use Questionnaire](#)
 - e. [Drug Screening Questionnaire \(DAST\)](#)
 - f. [Momentum Readership Questionnaire](#)
3. Look through each of the questionnaires above and see if there are any additional recommendations you would add to the list that you created.

Additional Readings

- Toonurangeau & Smith. (1996). ["Asking Sensitive Questions: The Impact Of Data Collection Mode, Question Format, And Question Context."](#)
- Krosnick & Alwin. (1987). ["An evaluation of cognitive theory of response-order effects in survey measurement."](#)

A close-up photograph of a person's hands holding a white smartphone. The person is sitting at a wooden table in what appears to be a cafe or restaurant. On the table, there is a white bowl of food, possibly a salad or soup, and a tall glass of orange juice with a pink straw. The background is blurred, showing other tables and chairs.

Part 4

How do I collect data?

Source: <http://www.qsffood.com/how-online-survey-can-help-you-to-improvise-your-sales-plan/>

Distribution

Once your survey is finalized, you next have to think about how you will distribute your survey. One of the first questions in this step is to consider when you will distribute it – particularly with regards to holidays, political cycles, life events, vacations, and so on. You can also think about who else is circulating surveys to your target population and consider whether your surveys are sent out to close to one another. Participants of some populations can become fatigued by the number of surveys they are asked to complete – and therefore opt out of completing them altogether. The second question you'll need to consider is how you are administering surveys, are you using in-person interviews? Computer assisted self-interviews (i.e., web or app-based surveys)? iPhone Surveys? Mailers? Each of these are good options with benefits and limitations. Which one you choose will be closely linked to how you are recruiting participants, who your target population is, and what information you are hoping to collect.

- During **in-person surveys**, researchers interact directly with participants and often ask questions verbally. Alternatively, they might just hand out a paper survey or a tablet on which participants can complete the survey (See CASI, below). Interviewers can extend person invites to individuals at a venue, or go door-to-door. These surveys tend to have higher response rates, decrease non-

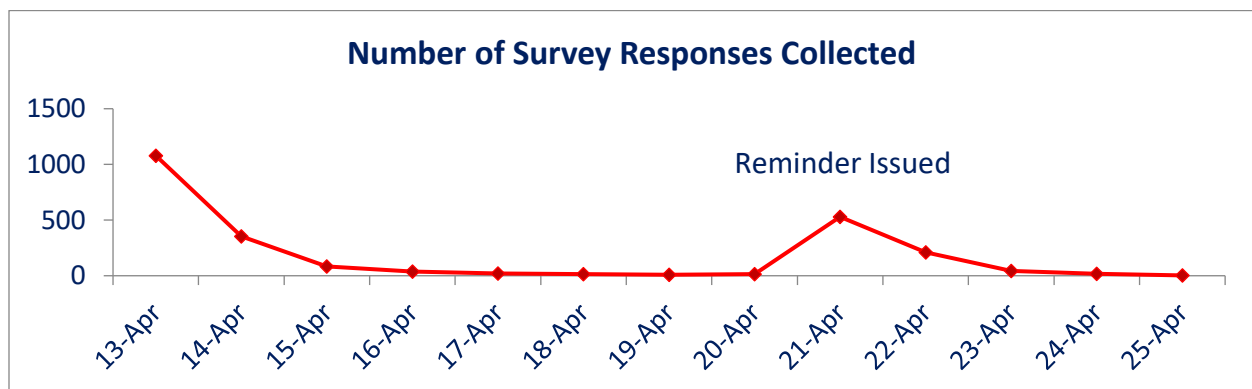
response, can be used to access hard-to reach populations (e.g., seniors), can be done in a variety of settings, allow researchers to make observations to enrich survey data, can involve all 5 senses (e.g., taste testing), and can be used to collect biological data that can accompany survey data (e.g., dried blood spot testing). Unfortunately, they tend to be more expensive, time consuming, and logistically challenging; and of course, interviewer biases can impact how participants respond to surveys.

- **Telephone surveys** are similar to in person surveys, but may be managed either digitally (using recordings and touch tone keypad responses) or manually. These also have better response rates, but are lower cost and less time-intensive than in-person surveys. As with in-person surveys, some populations are not accessible via this method (e.g., unlisted numbers, demographic differences). Further, telemarketers have given phone surveys a bad rap – introducing challenges in response bias.
- **Mailers** can also be used to reach participants with physical mailing addresses. Usually, mailers are accompanied by a letter of explanation, a self-addressed pre-paid envelope for returning the questionnaire, and the surveys themselves. In some cases, an honorarium may be pre-sent along with the survey to encourage participants to participate. Follow-up reminders (usually 2 follow-ups, 2-3 weeks apart) can be sent as postcards to encourage individuals to complete the survey. Mailers allow for larger samples, are cheaper than interviews, and allow respondents to complete the survey on their own time. However, they are more expensive than doing so online, and result in low response rates.
- **Computer-assisted self-interviews (CASI)** are like paper surveys, except they have been digitally retreated. Sometimes they are shared through a web-link or by downloading an app. Often times, participants are recruited on social media or from email list servers. CASI allows for a lot of customization and easily manages skip and display logic. Anonymous links can allow for sensitive data to be collected anonymously. However, if you are offering an incentive, you may want to ensure that participants are not responding more than once. If you have a set group of participants whom you might want to administer multiple surveys to you can create a panel of respondents. Panels allow you to track responses and send reminders for completion to those who have not yet

completed the survey. If reminders are sent, they are often sent after about 48 hours – but you can be flexible in choosing the time between reminders. In any case, data collection is in-expensive and automatically entered into a database. Further, you can force responses to certain questions. Of course, these surveys also have lower response rates and require participants to have a computer. Further, technical errors can make responding to surveys difficult.

As discussed in each of these examples a major consideration is the response rate of your survey. Whenever possible it is a good idea to measure the response rate of your survey. This can be done by dividing the number of people who were invited to complete your survey by the number of people who actually completed your survey. If possible it is also good to identify if there are differences between those who respond to your survey and those who do not. This is partially possible using ads-based delivery available on social media sites. High response rates are desirable because they improve the representativeness of your sample, help you get more diverse opinions, protect against nonresponse bias. When response rates are low, it may introduce bias and confounding if there are important differences in who chooses to respond.

Response rates generally fluctuate with between 20-35% of respondents completing a 10-minute survey. Higher response is influenced by the motivation of participants. Altruism and other intrinsic motivators can be leveraged by distributing the survey through trusted community leaders. Financial incentives can also encourage participation. Regarding incentives, raffles are generally not as useful as guaranteed small incentives – but don't be afraid to be creative in choosing how you motivate individuals to participate. Note that longer surveys tend to have lower response rates – particularly for individual questions that appear towards the end of the survey. As noted throughout this section, reminders can also be used to increase response rates.



Learning Activities

1. If you were conducting a survey among senior citizens what distribution method would you choose? Would you change the survey type if you were looking at the spread of sexually transmitted infections in this population? What about if you choose to look at young adults instead? Explain each answer.
2. Estimate the cost of conducting each of the following surveys for 1,000 participants:
 - a. Phone survey
 - b. In-person survey outside of a shopping mall
 - c. Online survey using survey monkey
3. What factors might motivate you to choose a more expensive survey design over a cheaper design?
4. Criminalization and stigma of behaviours makes it difficult to recruit participants and study these topics. How do you think you can overcome these barriers in the recruitment method?
5. Would you use different recruitment methods for doctors and patients if you were conducting an evaluation of attitudes towards a new therapy? If so, what elements might differ in these two approaches (e.g., venue, distribution method, incentives)?

Additional Readings

- Schillewaert & Meulemeester. (2005). “Comparing response distributions of offline and online data collection methods.”
- Christensen et al. (2017). “Cost and Efficiency of Online and Offline Recruitment Methods: A Web-Based Cohort Study.”
- Temple & Brown. (2011). “A Comparison of Internet-based Participant Recruitment Methods: Engaging the Hidden Population of Cannabis Users in Research.”
- Singer et al. (1999). “The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys.”

Chapter 6

QUANTITATIVE DATA ANALYSIS

Learning Objectives

- Plan analyses of quantitative data.
- Interpret quantitative results.



Part 1

How do I analyze quantitative data?

Source: <https://www.komando.com/tips/397615/ways-youre-ruining-your-computer-without-realizing-it/all>

Preparing Your Data

Once you have your data – meaning you entered it into an excel sheet or downloaded it from a survey collector – the first thing you will need to do is clean your data. Data cleaning involves removing duplicate entries, removing people who clicked through the survey but provided no usable data, recoding variables into the categories you want, and dealing with extreme values or erroneous data.

As you begin to edit your data it is important that you make a copy of your original data to work from. You should never over-write your original data. You will never know when you will need it. You should also make sure that each question is represented by a single column and that no column contains multiple pieces of information. The levels in each column should match the levels you want to use in your analysis, plus any extra levels you created to account for skip/display logic, missing data, and so on. Your final data set should have no blank cells (use some indicator, such as 9999, to indicate true missingness and other indicators to indicate why observations are missing).

Most survey data will have missing data somewhere. Respondents may not answer a question (either voluntarily or accidentally) or the question may not have been asked. It is important to distinguish between these types of missingness. Participants might

also give poor quality data – which you can recode as missing or as “Poor Data Quality.” This is common in text questions. For examples you might ask how many hours of community service somebody has done in the past year and they would report 10,000,000. Clearly this value is too extreme to be taken seriously. Once you have identified poor quality and missing data, you have a few different options:

- You can delete the entire row of data – a method called **list-wise** deletion.
- You can **leave it** missing and report the total number of responses separately for each variable.
- You can **impute** or assign some value (average, median, modeled value).
- You can **randomly assign** a value noting that this adds to random error.

Descriptive Statistics

Once your data is clean you will want to begin to describe your data. For numeric data you should consider where the distribution has its peak (*central location*), how widely dispersed it is on both sides of the peak (*spread*), how tightly the data are dispersed near the peak (kurtosis), and whether the distribution is symmetrically distributed or skewed. You can plot these data as a histogram and assess each thing visually. There are also more formal measures that are used that we call “summary statistics.”

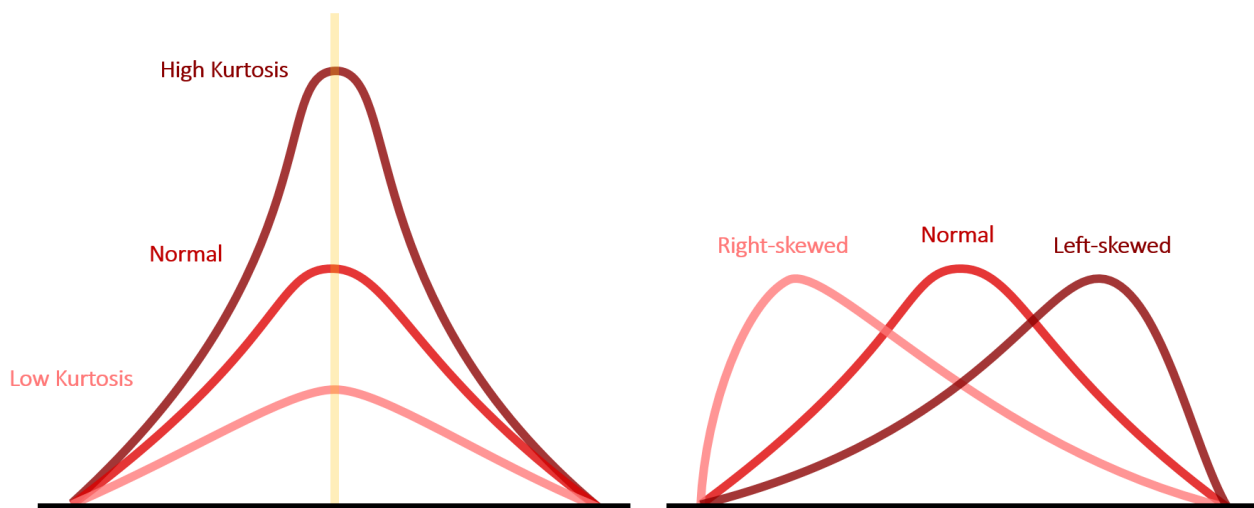
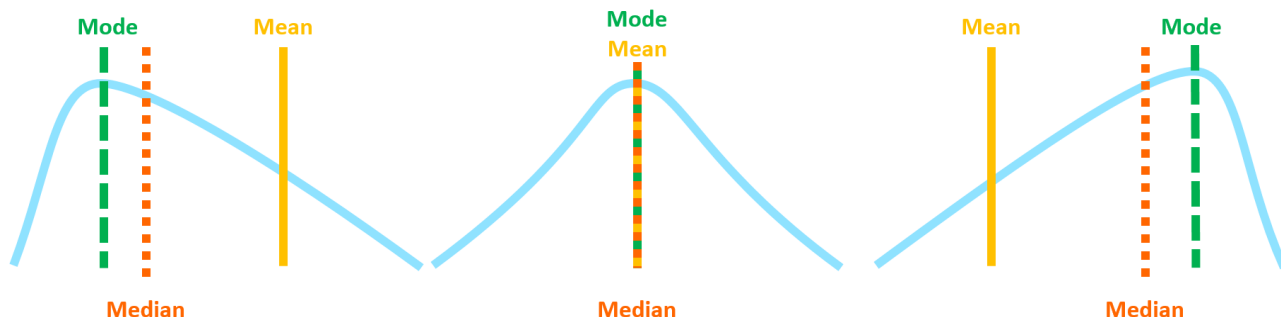


Figure 1. Kurtosis

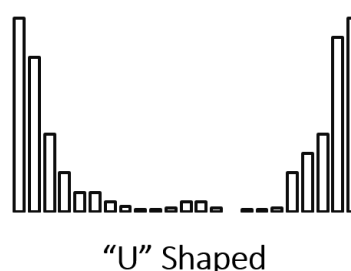
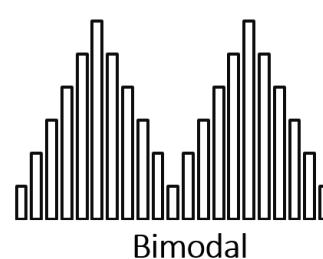
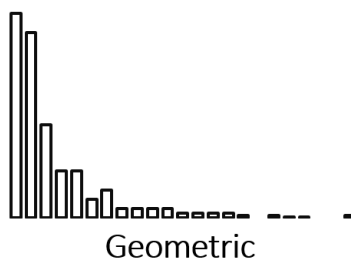
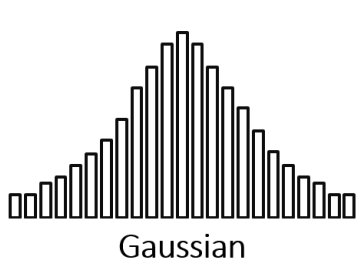
Figure 2. Skewness

Summary statistics for **measures of central location** include the mean, the median, and the mode. The mean is the average value of a dataset, the median is the middle value, and the mode is the most common value. If data are skewed, the median is a preferred

measure of central location, if the data is normally distributed the mean is often reported. It is good to calculate each summary statistic for each numeric variable in your dataset.



Creating a histogram of your numeric variables can be very informative. It can help you understand whether your data are normally distributed (i.e., have a Gaussian distribution) or whether they have some other distribution. Depending on the distribution of your data, you may or may not be able to use traditional statistical approaches. You should be aware that many measures in health – particularly measures that count things (e.g., number of days in hospital) – tend to have a distribution that is highly right skewed (e.g., Poisson, binomial, negative binomial, geometric). If your data is not normal, it is wise to consult a statistician prior to conducting your analyses.



Poisson
Binomial
Inverted U
Negative Binomial
Bernoulli
etc. etc.

In addition to measures of central location, measures of spread can help you analyze your data. Common measures of central location include the range, the variance, the standard deviation, and the interquartile range. You should recall the definitions of these from your biostatistics training. Boxplots can also be used to help you visually inspect the spread of your data. It is good to construct box plots and to calculate these statistics for each numeric variable.

Descriptive variables are usually much easier to describe. Traditionally we report the number of individuals in our sample who answered each question, and the number and the proportion of those answering the question that answered the question each way. For example, if you were reporting ethnicity, you would give the total number of respondents in your sample who provided their ethnicity and then report the number and percent of those who chose each response option. Bar charts are a good way to visualize this data.

Bivariable Statistics

In addition to describing your data, you will also likely want to conduct statistical tests to assess whether there is a statically significant association between two variables. In doing so, you will need to identify whether you are working with numeric or descriptive dependent and independent variables. Numeric variables include factors such as height, weight, BMI, and so fourth. Descriptive variables include factors such as ethnicity and HIV-status. If you are working with numeric data you will need to know whether your data is normally distributed and if you care comparing two or more groups, you will need to know if the variances are equal between them. Two important tests can help you assess normality and the equality of variances:

- The *Shapiro-Wick test* is used to assess normality of a numeric variable.
- The *Levene test* is used to test whether the variance of a numeric variable is equal between two groups.

Table 22 shows you which tests you can use given the specific data you are working with. You can also use [The Decision Tree for Statistics](#). As this is not an epidemiology or biostatistics course we will not get into all of the many nuances with these measures. Suffice it to say that you should know when to use these tests. In the real world you

will probably need to become more familiar with these or higher a statistician to assist you in your evaluation.

Table 22. Statistical Tests

Independent Variables (IV)	Independent Groups	Normality	Variances	Dependent Variable (DV)	Test(s)
Correlation between two numeric variables					
Numeric	N/A	Normal	N/A	Numeric	Pearson's Correlation
Numeric	N/A	Non-Normal	N/A	Numeric	Spearman's Rank or Kendall's Rank Correlation
Association between 2 Groups					
Numeric	Independent	Normal	Equal	Descriptive	Student's t-test or Linear Regression
Numeric	Independent	Normal	Not Equal	Descriptive	Welch's t-test
Numeric	Paired	Normal	Equal	Descriptive	Paired t-test
Numeric	Independent	Non-Normal	N/A	Descriptive	Mann-Whitney U test
Numeric	Paired	Non-Normal	N/A	Descriptive	Wilcoxon Signed Rank test
Descriptive	Independent	N/A	N/A	Descriptive	X ² test
Descriptive	Paired	N/A	N/A	Descriptive	McNemar's test
Descriptive – Two Levels	Independent	N/A	N/A	Descriptive	Binomial Logistic Regression
Descriptive – Two Levels	Independent	N/A	N/A	Numeric	Binomial Logistic Regression
Association between >2 Groups					
Numeric	Independent	Normal	Equal	Descriptive	One-way ANOVA
Numeric	Independent	Normal	Not Equal	Descriptive	Welch's One-way ANOVA
Numeric	Independent	Non-Normal	N/A	Descriptive	Kruskal Wallis H test
Descriptive – More than Two Levels	Independent	N/A	N/A	Descriptive	Multinomial Logistic Regression
Descriptive – More than Two Levels	Independent	N/A	N/A	Numeric	Multinomial Logistic Regression

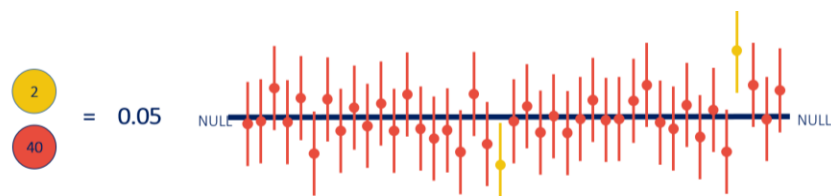
In addition to knowing when to apply various tests, it is also important that you are able to provide accurate interpretations of the results from these tests. I find that many tests are easily interpreted incorrectly. Therefore, let's cover some of the basic interpretations for the output of the most common statistical measures:

- The **linear slope** from a linear regression model is interpreted as the unit increase in the dependent variable for each unit increase in the independent variable. Often only the p-value is interpreted and the slope is mistakenly ignored. Do not ignore the slope! The sign (+/-) of the slope indicates whether the association is significantly positive or negative.
- The **Pearson's r** is the strength and direction of an association. Values closer to 0.0 indicate a weak association, those around 0.7 indicate a moderate association, and those near 1.0 indicate a strong relationship. The sign (+/-) of the Pearson's r indicates whether the association is significantly positive or negative.
- The **R²** is the degree to which the observed data fit the modeled line. The R² can only have positive values. The closer the value is to 1.0, the stronger the data fits the line; the closer the value is to 0.0, the less the data fits the line.
- **Odds ratios** are used to compare the relative effect of one group to the "reference" group. Values greater than 1.00 indicate a positive association and values less than 1.00 indicate a negative association.
- **Relative risk** is interpreted the same way as an odds ratio, except it represents risk and not odds. To calculate risk, you need to know the number of people at risk for a disease. This is usually not possible in case-control studies, but is possible in cohort studies. Odds and risks can sometimes be interpreted the same way, however, doing so tends to inflate the association. Take for example a hypothetical group of 129 seeking treatment for bleeding esophageal varices. Imagined that 65 people received endoscopic sclerotherapy and 64 received band ligation. Of those receiving band ligation (n = 64), 18 died. Of those receiving sclerotherapy (n = 65), 29 died. The overall risk of death in this scenario would be 0.36 (n = 47 died out of 129) while the overall odds of death would be 0.57 (n = 47 died and 87 lived). You can see that if odds are interpreted as risk, the interpretation would give you an inflated estimate of risk.
- **P-values.** To understand how to interpret the p-value it is helpful to know how not to interpret it: The p-value is not the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false; The p-value is not the probability that the observed effects were produced by random chance alone; and the p-value does not indicate the size or importance of the observed effect. A very "significant" p-value could correlate with a minuscule effect. The 0.05 significance level is merely a convention and is nearly indistinguishable from 0.04 or 0.06; yet often times

statements regarding “significance” are made based on threshold significance. The p-value is the probability that you are committing a type 1 error. A type 1 error is a false positive – the rejection of a true null hypothesis. The probability of obtaining data as extreme, or more extreme, than those observed if the null hypothesis is correct.

- **Confidence Intervals.** A 95% CI simply means that if the study is conducted multiple times (multiple sampling from the same population) with corresponding 95% CI for the mean constructed, we expect 95% of these CIs to contain the true population mean. A common misunderstanding about CIs is that with a 95% CI there is a 95% probability that the true population mean lies between A and B. This is an incorrect interpretation of 95% CI because the true population mean is a fixed unknown value that is either inside or outside the CI with 100% certainty. In other words, the inclusion of a true population mean is not a probabilistic occurrence. Furthermore, remember that like with the p-value, the choice of whether to use a 90% or 95% CI is somewhat arbitrary, and depends on the level of “confidence” that the investigator wishes to convey in his or her estimate.

When interpreting statistical significance, it is important to look at all the comparisons being done in a study. This is because as the number of tests performed increases, the family-wise error rate (FWER) also increases. The FWER is the probability of obtaining at least one false positive when the null hypothesis is true. In other words, if you are choosing an alpha level of 0.05, then you are saying that out of 40 tests, you anticipate at least 2 would occur by chance alone. This reality has given rise to a phenomena called p-hacking in which the tests being conducted are repeated until a spurious finding can be identified and reported on. Given what you know about the FWER, be careful in interpreting p-values that are close to your cut off levels and avoid reporting p-values as simply “significant” or “non-significant.” It is better to communicate the meaning of a p-value not just its relevant to your hypothesis test.



Finally, you should be careful to make sure that you are accounting for confounding and bias in your analysis. If you did not use randomization to assign participants to non-

participants, multiple regression and other statistical approaches can be used to control for confounding. It is recommended that you work with a statistician on these more advanced approaches. When conducting any of the analyses discussed in this chapter, you can calculate them by hand or use statistical software such as SPSS, SAS, STATA, or R. I generally recommend R for situations in which you will not have access to the other expensive software. Furthermore, R is open source and more adaptable and its algorithms have been peer reviewed. I recommend that you take the time to learn R if that is at all possible. For those of you who will go on to use R, the table below provides a helpful overview of the syntax that is used to create the analysis discussed in this chapter:

Table 23. Common Functions in R

Need	Function
Read in your dataset.	<code>read.csv("<path location of your csv file>")</code>
Subletting your dataset into a smaller dataset	<code><new data> <- [<old data>[which(<old data>\$<old variable> == "<desired group to include in new dataset>"),]</code>
Identify whether a variable is numeric or factor.	<code>class(<data>\$<variable>)</code>
Identify all the levels in a variable.	<code>levels(<data>\$<factor variable>)</code>
Relevel a variable	<code><data>\$<new variable>[<data>\$<old variable> == "<old level>"] <- "<new level 1>"</code> <code><data>\$<new variable>[<data>\$<old variable> == "<old level>"] <- "<new level ...>"</code> <code><data>\$<new variable>[<data>\$<old variable> == "<old level>"] <- "<new level x>"</code>
Convert a categorical factor to a numeric variable.	<code><data>\$<factor variable> <- as.numeric(<data>\$<numeric variable>)</code>
Convert a numeric variable to a categorical factor.	<code><data>\$<numeric variable> <- as.factor(<data>\$<numeric variable>)</code>
Calculate the mean, median, quartile, minimum, or maximum values of a continuous variable.	<code>summary(<data>\$<numeric variable>)</code>
Calculate the standard deviation of a continuous variable.	<code>sd(<data>\$<numeric variable>)</code>
Calculate numeric statistics separately for multiple groups.	<code>by(data = (<data>\$<numeric variable>, INDICES = (<data>\$<grouping factor>, FUN = summary)</code>
Calculate the frequency of responses across a categorical factor variable.	<code>table(<data>\$<factor variable>)</code>
Calculate the proportion of responses for each level of a categorical factor variable	<code>prop.table(table(<data>\$<factor variable>))</code>
Create a cross tabulation of two categorical factor variables	<code>table(<data>\$<factor variable>, <data>\$<factor variable>)</code>
Create cross tabulated proportions of two categorical factor variables	<code>prop.table(table (<data>\$<factor variable>, <data>\$<factor variable>), margin = 1)</code>

Note: words within brackets "<>" indicate text to be filled and are not the code syntax.

Learning Activities

1. Provide an interpretation for each of the statistics testing the relationship between age and sexual debut (age at first sexual encounter).
 - a. $R^2 = 0.92$
 - b. $p = 0.06$
 - c. $p = 0.001$
 - d. $r = 0.78$
 - e. Linear Estimate = 0.71, 95% CI = 0.11, 1.52
2. If a median is twice the value of your mean, what does this indicate about the distribution of a numeric variable?
3. Why would you use the Shapiro or Levine tests?
4. If you were testing the association between the onset of cardiovascular disease (yes vs. no) in adulthood and childhood depression scores (range: 1 to 30), what test would you use?
5. What is an equivalent graph or figure to the histogram for descriptive (categorical) data?

Additional Readings

- Poole (2001). "Low P-values or Narrow Confidence Intervals: Which are More Durable?"
- Greenland et al. (2016). "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations."
- German & Loken. (2013). "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time."
- The Decision Tree for Statistics

QUALITATIVE DATA COLLECTION

Learning Objectives

- Design a qualitative study.
- Develop an interview guide for one-on-one interviews or for a focus group.

Part 1

How do I collect qualitative data?



Source: <http://masteremergencyarchitecture.com/tag/slums/>

Importance of Qualitative Approaches

In many situations, quantitative data is not feasible or appropriate for answering a specific research question. This is particularly so, when not much is known about a topic or when the results of a quantitative study do not actually tell you much about the phenomena of interest. The use of qualitative methods can tell you about lived experiences, meanings, and perspectives and can be used to probe the depths of an issue in a way that is not possible with quantitative data. While some view qualitative methods with suspicion, this worldview is not justified. Qualitative methods are important and necessary for evaluations. The goals of qualitative research can be classified using four general aims:

- **Exploration**, which includes discovering themes and patterns to build an initial understanding of a complex phenomenon, asks questions such as “What kind of things are here or going on?” “How are these things related to one another?” “Do these things fall into natural groups or categories?”
- **Description**, which includes providing an illustration of a phenomenon, asks questions such as “What does this look like?” “Why does it happen?” “What is its purpose?” “Who is involved?” “What are their roles?”

- **Comparison**, which includes looking for differences and similarities in things, asks questions such as “How does X differ from Y?” “How is this group different?” “What factors might drive these differences?”
- **Modelling**, which includes testing conceptualizations developed in the previous steps (or through a priori experience) against observations of new data, asks questions such as “Does this case conform to my general model?” “If not, how is this case different and how should I incorporate it in my model?”

Types of Qualitative Studies

In quantitative methods, there are five basic approaches that can be used to achieve the aims described above: interviews, focus groups, document reviews, participatory methods, and observations. As we already covered observations in an earlier chapter, we will discuss each of the other four methods here:

Interviews with Individual Respondents

Under the umbrella of one-on-one interviews exists a gradation of interviewing styles ranging from informal conversations to structured interviews. These interviewing styles may be used to explore ethnographic data (i.e., study of people’s experiences) or to illicit insight or opinions about a phenomena.

- **Informal conversations.** Informal conversations usually are not planned in great detail. While you may have a general idea of what you want to talk about and who you want to ask, informal conversations allow for greater flexibility. These conversations are generally not transcribed word for word – though you might take notes. The conversational nature of these interviews allows interviewers and interviewees to navigate through issues with less social formality. Another important characteristic of these informal interviews is the rapport upon which they are built. While other interview styles may take place between complete strangers, conversational interviews often occur when the evaluator and participant are embedded within the evaluation process together. The dynamic of interviewer-interviewee is thus de-emphasized and the flow of the conversation is less “question-answer” and more “how about this, how about that.”

- **Semi-structured interviews** take things one step farther by creating a somewhat more rigid framework for guiding the conversation. These interviews are often conducted between strangers and are often transcribed word for word. While the topics of various questions have been pre-established and formalized in an ethics guide, the interviews are generally given significant flexibility in how questions are asked, the order in which questions are asked, and the degree to which statements from the respondent are fleshed out. That said, interviewers should still avoid leading questions and design questions that illicit lengthy descriptive answers. The questions should be concise and easily understood as to avoid having to restate the questions. When ordering questions, you generally begin with less sensitive more general questions, but make sure you end on a positive note that's not too emotionally "heavy." If you are studying a process of life course, it is good to start with earlier events and move on to more recent events. If participants give shorter answers than expected have a few back-up questions and probing questions ready. When participants go in greater depth and answer questions that are intended to appear later – interviewers should adapt by not asking the questions again. That said, an interviewer also needs to be skilled in redirecting discussion back to the topic at hand. While tangents can lead to interesting information – it is important to keep the interview at a reasonable time limit.
- **Structured/standardized open-ended interviews** are the most rigid of the interview styles discussed here. They can technically be administered in person or by using open ended questions as part of a survey. These interviews are very rigid. The questions are asked the same way for all participants and in the same order. Probing questions, if used, are pre-planned as part of the questionnaire. These are sometimes used because a validated scale or questionnaire requires a specific structure to be accepted as valid. When the interviewer is not an area expert, such as during telephone interviews, these types of surveys can be useful as they do not require an agile and adept individual to conduct the interview.

Group Interviews

As with individual interviews, group interviews can be equally flexible or rigid. Nevertheless, they tend to be more similar to semi-structured interviews than anything else. This is because the interviewer will need to allow for the dynamics introduced when multiple people are being interviewed at the same time.

An important distinction between individual interviews and group interviews is that data from focus groups does not represent the viewpoint of an individual. Rather, focus groups will often represent the views of the most vocal individuals or the views of individuals who are delivering answers for the group. This can introduce bias, but can also be considered a strength of the group interview design since it tells you a bit about how individuals express their view points in inherently complex social settings.

Given that individuals are not the unit of analysis in a group interview, the questions and research questions generally focus on eliciting diverse viewpoints. The voices of various experts can help you reach new depths by identifying areas that a single participant might not recognize on their own. Sometimes focus groups can be used to identify a consensus opinion or perhaps identify the key issues of contention. Focus groups are often audio recorded and transcribed. Notes may also be kept by a second interviewer/observer. Body language and other non-verbal speech can also be examined and documented as part of the focus group.

Focus group interviews can be expensive and require multiple rounds to get all the issues at hand. Generally speaking, 5-7 people participate in a focus group. However, as many as 10 individuals might participate. The number of participants should be based on the expertise of the participants. In considering the composition of focus group, interviewers should consider power dynamics and how these can limit some individuals from sharing their perspectives. Focus groups also tend to be a bit longer than one-on-one interviews. This allows greater involvement from the larger number of participants. Typically, 2-3 hours is the maximum duration for a focus group, erring towards 2 hours

“Document” Reviews

Not all qualitative work requires interviews with participants. Legal decisions, policy documents, staff communications, videos, music, news media articles, and social media posts all provide a great source of qualitative data. These data sources can be used to

understand the discourse of a debate or to understand how a particular program, policy, or intervention is viewed on the mass scale. Increasingly, natural language processing and sentiment analysis techniques can be used to understand the relationship between written concepts. Comparisons of how documents change over time or how representations of ideas over time can be extremely useful in understanding shifting opinions and viewpoints. In many ways, document reviews allow you to skip data collection and transcription and go straight to analysis. Given this, a document review can be a cheap and effective way to get much of the same information that might come from personal interviews.

Participatory Methods

On the opposite end of the spectrum, participatory methods involve a much greater level of engagement from participants – often involving them in the knowledge generation and analysis processes. There are many great examples of participatory approaches – including those covered in the Appendix. Of the various qualitative data collection methods, participatory research methods are the most rapidly expanding and least well defined methodologies. This is largely because of the diverse engagement strategies that are employed. Body mapping, for example, asks participants to map their bodies creatively and walk the interviewer through the meaning they ascribe to their art. Other methods might entail a participant giving an interviewer a tour of their apartment or work place or creating a video about some of their daily experiences.

These methods are as varied as the research questions they aim to address. The key unifying aspect is the participatory nature that engages participants and gives them power to express themselves as they want to be perceived. This helps prevent boredom and fatigue among participants and may relax individuals who are not accustomed to being interviewed. It is widely used for interviewing populations for whom traditional interviewing styles might not be appropriate or feasible (e.g., youth).

“Interview” Guides

Regardless of which qualitative techniques are selected, it is important that the process of data collection is well-documented. This helps you to account for sources of bias, including interviewer biases that can emerge from inconsistencies in the ways data are collected. Interviewer guides are not only important for when multiple interviewers are engaged in the interview process, but also for when those collecting the data are not

the ones who conducted each of the interviews. Among the various things an interview guide establishes, it identifies the main questions and probing questions that are used to flesh out details or information. Probing is key to successful interviewing as it allows for a more in-depth examination of an issue. There are a variety of probing type questions that can be used:

- **The silent probe** involves waiting for participants to continue talking after they have stopped. It is a small signal to the speaker that they should continue speaking.
- **The echo probe** involves simply repeating back to participants what they last said. It shows that you understand what they said and allows them to add more information if they want to.
- **The Uh-huh probe** is the use of short phrases or sounds to confirm to the speaker that you are engaged and listening. It can help them continue in the current line of thought.
- **The tell-me-more probe** involves asking things like “Can you tell me more about that?” or “Why do you say that?” or “How does that make you feel?”
- **The long question probe** is the use of longer questions to illicit longer responses. For example instead of saying “Why did you start using methamphetamine?” you might say “Some people start using methamphetamine because they are trying to cope with life or to build social connection. Why did you start using methamphetamines?” These sorts of questions are good for sensitive topics because they create comfort and help the hearer to understand where you’re coming from.

So, while interviewer guides should include descriptions of the primary and probing questions to be used, they should go beyond this and also identify the

- **characteristics of the interviewer/facilitator** (i.e., credentials, occupation, gender, experience, relationship with participants, how they establish relationships, participant knowledge of interviewer, and other important characteristics);
- **theoretical framework** guiding the development and interpretation of questions;
- **recruitment methods** (i.e., sampling, mode of interview, sample size, response rate);
- **setting** (i.e., presence of non-participants, location and style of interview room);

- **data collection tools** (i.e., number of interviewers, recording, note taking, duration; and transcription methods); and
- information about **how data will be analyzed** (i.e., number of coders, description of codebook; description of how themes will be identified; description of software used in coding process; description of how participants will review findings; description of how quotes will be selected).

All of this information is a helpful part of the interview guide because it helps situate the discrete experience of interviewing into the broader qualitative research framework. Thinking through these aspects will ensure that you have the information you need to successfully execute your study and report on your findings. Further, having information about the interviewer, environment, and other key issues before the interviewer begins will help you to better account for these factors. For example, research has shown that *the deference effect* – which is the effect that differences between the interviewer and participant have on a participant’s responses – can play a significant role in shaping how an interview unfolds. By thinking through this issue ahead of time you can better identify strategies to mitigate this during the interview.

As you think about what should be considered in the lead up to an interview or focus group, two helpful resources include (1) the consolidated criteria for reporting qualitative research (COREQ) and (2) the Evaluation Checklist Project’s checklist for guiding the development of qualitative research.

It is also important that in designing an interview guide that you consult stakeholders to ensure that they feel the questions being asked are within the scope of your project. Stakeholders with lived experience dealing with the issue at hand will help you flesh out probing questions and identify areas that you might have otherwise ignored. Further, involving stakeholders early in the study design process will encourage buy-in and investment. Invested stakeholders may be more willing to participate in the interpretation of data and in the data analysis process. Having these extra helpers can be of great benefit to you throughout the qualitative research process.

Learning Activities

1. What are the limitations of focus groups?
2. Imagine you are evaluating the implementation of a safe consumption site. What research questions might you investigate using qualitative methods? Which of these research questions would be best suited for focus groups and which would be best suited for one-on-one interviews.
3. In qualitative research you are less concerned with generalizability and more concerned with saturation. Saturation is the concept that data should be collected until new data or analysis of data is no longer necessary (i.e., additional data would not add anything new). Describe how you might consider saturation as you plan, conduct, and evaluate your qualitative research.
4. What do participatory evaluation methods offer over traditional qualitative interviews and focus groups?
5. True or False: Structured Interviews can be conducted over the phone, in a survey, or in person.
6. Why do you think it is important to report information about the interviewer and interview setting?

Additional Readings

- Thackeray & Geiger (2004). "[Misconceptions of Focus Groups: Implications for Health Education Practice.](#)"
- Tong, Sainsbury, & Craig. (2007). "[Consolidated criteria for reporting qualitative research \(COREQ\): a 32-item checklist for interviews and focus groups.](#)"
- Patton (2003). "[Qualitative Evaluation Checklist.](#)" Evaluation Checklist Project.
- Saunders et al. (2018). "[Saturation in qualitative research: Exploring its conceptualization and operationalization.](#)"
- [Example of Semi-Structured Interview Guide](#)
- [Example Structured Interview Guide](#)
- [Example Focus Group Discussion Guide](#)

QUALITATIVE DATA ANALYSIS

Learning Objectives

- Plan analyses of qualitative data.
- Interpret qualitative results.



Part 1

How do I analyze qualitative data?

Source: <https://www.achievable.co.uk/evasys/how-to-effectively-carry-out-a-qualitative-data-analysis>

The Process of Analyzing Qualitative Data

Quantitative data analysis involves five primary steps: (1) organizing data, (2) finding and organizing concepts, (3) building overarching themes, (4) ensuring reliability and validity, and (5) rationalizing findings.

Organizing Data

Data from qualitative studies are first transcribed into a format in which they can be analyzed. According to Bailey (2008), “Transcribing appears to be a straightforward technical task, but in fact involves judgements about what level of detail to choose (e.g. omitting non-verbal dimensions of interaction), data interpretation (e.g. distinguishing 'I don't, no' from 'I don't know') and data representation (e.g. representing the verbalization 'hwarryuhh' as 'How are you?'). Representation of audible and visual data into written form is an interpretive process which is therefore the first step in analysing data. Different levels of detail and different representations of data will be required for projects with differing aims and methodological approaches.”

Once data is transcribed it can be analyzed in specialized software (e.g., NVIVO, ATALS.ti, QDA Miner, Tams Analyzer, Dedoose, MAXQDA, HyperRESEARCH, AQUAD, Mendeley,

and Transana). These softwares, to varying extents, facilitate the remaining steps in the qualitative data analysis process.

Organizing Concepts & Building Themes

Data from qualitative interviews, focus groups, or any other source can be analyzed using primarily two types of qualitative methods: deductive methods and inductive methods.

- **Deductive methods** are used when a pre-existing theory or framework is used to explore the data. Pre-existing themes and codes, including entire codebooks, can be used to study the data.
- **Inductive methods** are used when there is a lack of previous theory or findings on a topic or when the authors choose to disregard existing theory for the sake of evaluating data from a fresh perspective.

Building off these two general methodological approaches, there are a variety of specific analytic frameworks that can be applied to qualitative data. There are undoubtedly many of these in existence, but the five listed here capture the bulk of qualitative methodology.

- **Content analyses** are either inductive or deductive approaches that examine the content of a data source – often with the goal of quantifying trends and patterns of words used, their frequencies, their relationships, and the structures and discourses of communication.
- **Thematic analyses** are either inductive or deductive approaches that examine the data in order to identify key themes or ideas that are represented by the content of a data source.
- **Discourse analyses** are either inductive or deductive approaches that examine naturally occurring talk and all types of written text to understand how language is used.
- **Narrative analyses** are either inductive or deductive approaches that examine the stories and narratives shared by participants.
- **Grounded analyses** are primarily deductive approaches that examine data sources without respect to pre-existing theory.
- **Phenomenological analyses** are primarily a deductive approach that attempts to describe a phenomenon by articulating its essential nature.

With each of the analytic frameworks listed above, you will need to begin your analysis by creating codes – a process referred to as coding. Codes are themes or short words or phrases that represent a key theme or idea. There are two types of coding used in the coding process:

- **Open coding** involves the organization of raw data to try to understand what has been captured.
- **Axial coding** involves connecting and linking the codes together.
- **Selective coding** involves identifying the core of the issue at hand.

Although coding styles vary between researchers, some of the main questions addressed in open coding include:

- What are the underlying issue and the phenomenon?
- Who are the actors involved and what are their roles?
- How, when, and where does a phenomenon take place?
- How intense or salient is the experience?
- Why and for what reasons does the phenomenon occur?
- Which strategies and tactics facilitate the phenomenon?

When more specific theoretical frameworks are being utilized, specific dimensions of the framework are included. It is important to recognize that in developing codes, you are working to identify both subtle and obvious expressions and themes. Codes can be identified as important based on (1) how frequently they appear, (2) how pervasive they are across different domains, (3) how people conceptualize or interact with the code, and (4) the degree to which the code is influenced by specific contexts or situations. When looking to identify a new code there are several things you can look for:

- Frequent repetitions and circling back to a core issue or element.
- The presence of unfamiliar word or words that are used in unfamiliar ways.
- The use of metaphors or analogies.
- Transitions in tone, thought, or demeanor during a conversation.
- Linguistic connectors that imply causality (e.g., because), conditionality (e.g., if-then), taxonomy (e.g., is a), temporality (e.g., before, after, next), and location (e.g., is close to, by, near).
- Unbalanced qualifiers (e.g., talk about health and women’s health, but not men’s health).

- What’s missing (e.g., is there something important or central that just doesn’t get mentioned.)

After the open coding process is complete, axial coding can be used to link specific ideas together; and selective coding can be used to identify the underlying theme that links all of the axial codes together. Again, these processes are highly specific to the exact methods chosen for your analysis and the theories, models, and frameworks that are used to interpret your data. That said, there are a few widely used techniques that can help you process your codes. Among these techniques are:

- **Cutting and sorting** traditionally involves marking each sentence line by line and then cutting each quote out and pasting it to a small index card. The index sorts are then piled based on how similar they are to one another and to what degree the codes relate to one another. Two strategies – splitting and lumping – can be used to either maximize the differences and create more refined themes or minimize the differences and create more general themes.
- **Word lists and key-words-in-context** strategies involve counting the number of times each words appear and the frequencies they appear in which contexts (e.g., Do pretty and handsome appear more frequently for one gender or the other).
- **Word co-occurrence** involves looking at how frequently certain words co-occur. For examine knowing that “Shrouded” occurs frequently with “mystery” or “secrecy” can give you a sense of a common theme. Likewise, crime may appear more frequently with violence than it does with justice – suggesting two potentially distinct themes: criminality (negative) and justice (positive).

Choosing a coding technique

When choosing a technique, it will depend on the kind of data you have (e.g., Is it text or something else? Do you have verbatim text or just field notes? Are the narratives rich or is your data mostly just short response?), the capacity of your team (e.g., Do you have enough cultural competence to spot metaphors, connections, and missingness?; Do you have the technical skills to construct co-occurrence data or to use other quantitative analysis methods?; Do you have the man hours to finish the analysis?); and how many

themes do you want to identify (cutting and sorting is flexible, but KWIC and word lists tend to produce a lot of themes while meta-coding and co-occurrence analysis produce fewer themes).

Codebooks

Once you have identified your themes, it is time to begin developing a codebook. Codebooks compile three types of codes: structural codes, theme codes, and memos.

- Structural codes provide information about the reviewer, environment, and interviewee.
- Theme codes capture the themes identified in the last step – these are the substance of what you are studying.
- Memos are field notes about the codes and contain our running commentary as we read through the text.

Codebooks can be developed deductively, but you can also have a predefined inductively developed codebook. For example, you might search the literature in a systematic review for all the psychological models of trauma and then create a codebook that allows you to identify cases in support of each model. Some fields or topic areas have pre-established code books that are available for use or that can be purchased.

Whether you choose to use or build a deductive or inductive code book is largely up to you and your stakeholders. Common features of codebooks included the anonymization of participants, the hierarchical organization of codes and their relationship to one another, and the instructions or criteria used to classify a participant response into the code.

When organizing codes hierarchically, it can be helpful to first identify the general codes and then identifying the sub codes. In some studies the general codes are identified first and then a separate process or analysis is undertaken to further highlight elements of each constituent code.

Ensuring Reliability and Validity

While reliability and validity are regularly considered in the course of quantitative work, recently there has emerged a growing interest in applying these concepts to qualitative

data. Reliability is probably the more emphasized element. Evaluators are often concerned with whether the individuals coding the data are coding the same passages the same way. Using multiple coders for each set of data or training the coders on a smaller test-batch of data can help improve the inter-rater reliability of your analysis. Training coders has all but become a standard approach for qualitative data analysis. Training consists of six steps:

- Give all coders a codebook they can use to code data.
- Review the codebook as a group and have coders code a set of real examples.
- Review their coding as a group and discuss and resolve discrepancies.
- Update the codebook as coders come to agreements about the content of each theme.
- Once coders start coding the bulk of the data, do random spot checking to ensure agreement.
- Repeat the steps above as often as needed to improve the reliability of codes. Generally agreement of around 80% is the minimum acceptable level for a reliably coded dataset. This means that in a random sample of 10 sections, 8 will have been coded the same way.

Validity is also increasingly considered in qualitative work. One of the main ways that validity of interpretations can be confirmed is through verification. Verification is the process of checking, confirming, and making sure that your results agree with the data and that the claims you make account for each observation. The process of rechecking your data against itself is sometimes called a “constant comparative” approach.

During the verification stage, you should systematically review your codes to ensure that coded sections are appropriately classified. Further, you can verify your data by interpreting the data as you collect it. This allows you to probe participants for information that will help you clarify your codes. Further, you can directly ask participants about key aspects of your emerging coding scheme to see if it has obvious face validity. Likewise, you can continue collecting data, until the addition of new respondents or experiences does not result in changes to how your data is coding – a condition referred to as saturation. Nobel & Smith ([2015](#)) provide a thorough list of additional checks to support validity and reliability:

- Accounting for personal biases which may have influenced findings;

- Acknowledging biases in sampling and data collection;
- Meticulous record keeping, demonstrating a clear decision trail and ensuring interpretations of data are consistent and transparent;
- Seeking out similarities and differences across accounts to ensure different perspectives are represented;
- Including rich and thick verbatim descriptions of participants' accounts to support findings;
- Demonstrating clarity in terms of thought processes during data analysis and subsequent interpretations;
- Engaging with other researchers to reduce research bias;

Incorporating these sorts of activities as explicit components of your study design will help support the conclusions of your qualitative study and show the rigour inherent in your approach.

Rationalizing Your Findings

It is important to realize that qualitative data analysis is an inherently cognitive process that requires you to understand a phenomenon, synthesize a portrait that accounts for all observed data, and theorizing about the relationships between phenomena. There are many more analytic approaches used in qualitative data analysis than are discussed here. As part of your qualitative data analysis process it is important that you consider your data and your analytic methods with respect to what is already known and assess to what extent your findings can explain these past observations. Doing so will help you to create a fuller picture of the phenomenon you are studying.

Learning Activities

1. Review one of the additional readings and create a concise list of steps needed to complete one of the analytical methods covered in your selected reading.
2. Under what conditions would you choose to take a deductive approach in a qualitative study?
3. Look at the trending categories on twitter and skim through the responses. Try creating three or four open codes that are connected to one larger axial code. Use the qualitative method and theoretical frame of your choice.
4. Conduct a google image search for “smoking prevention,” “drunk driving,” or some other topic of interest and create a code for each public health message being communicated in these pictures. Stop creating codes when you are confident that your coding scheme is saturated (i.e., looking at new pictures won’t change your results).
5. Based on the last activity, what are the axial and selective codes that link your observed codes together?

Additional Readings

- Erlingsson & Brysiewicz. (2017). “A hands-on guide to doing content analysis.”
- Braun and Clark. (2006). “Using thematic analysis in psychology.”
- Carbo et al. (2016). “How do I do Discourse Analysis?” Teaching Discourse Analysis to novice researchers through a study of intimate partner gender violence among migrant women.”
- Emden (1998). “Conduct a narrative analysis.”
- Strauss & Corbin (1994). “Grounded Theory Methodology: An Overview.”
- Beck (1992). “The lived experience of postpartum depression: A phenomenological study.”

Chapter 9

MIXED-METHODS

Learning Objectives

- Identify mixed methods study designs.
- Design a mixed methods study.



Part 1

What are mixed-methods?

Source: <https://www.improvediagnosis.org/researchresources/>

Mixed Method Research

Mixed-methods research involves utilizing both quantitative and qualitative methods. Creswell (2003) developed a typology for classifying mixed methods design. Sequential designs are those in which one data collection method follows the other. Concurrent designs are those in which both quantitative and qualitative data are collected at the same time. We will discuss four of these design types:

- The first sequential design is the **explanatory design**. In this design, quantitative data collection proceeds qualitative data collection. Qualitative results are used to explain the results of the quantitative study.
- The second sequential design is the **exploratory design**. In this design, quantitative data collection follows qualitative data collection. The qualitative data is used to explore an issue and help develop a quantitative instrument (e.g., scale, survey, etc.).
- The first concurrent design is the **triangulation design**. In this design, data are collected at the same time and used to cross-validate the findings of the other data source. It is very helpful when one design has some significant limitation that can be overcome by the other design. For example, a qualitative study might not be generalizable, but the results are supported with similar data from a quantitative study.

- The second concurrent design is the nested design. Often times a qualitative study is nested within a larger quantitative study. The aims of the qualitative sub-study may or may not relate to the methods of the quantitative study.

Regardless of which of these approaches you take, each design will have a “point of integration,” wherein the qualitative and quantitative components are brought together. There are six ways in which integration is commonly done:

- **Complimentary** methods can back-up the findings from other methods.
- Findings of one method can help **develop** the study design of another method.
- Comparing findings from each method for incongruences can help you identify key incites and **initiate** new lines of inquiry.
- By using different methods, you can **expand** your focus and address multiple issues at once.
- Data from multiple methods can help you **triangulate** results to address the limitations of one method by using results from the other method.
- Data from one method is **mixed** with data from the other method as part of the analysis of the data (e.g., qualitatively reported beliefs about a healthy service are analyzed separately for high income and low income respondents).

In thinking about mixed methods research, it is important to realize that one method is not simply done in service of the other. Rather, each method has its own unique strengths and contributions. Both quantitative and qualitative methods can help you to better understand cause and effect, understand nuanced social interactions, test hypotheses, and generate models and theories that describe how things work. Increasingly, evaluation and health research studies require the use of mixed methods. As such, the evaluation or research team will need to develop expertise, or at least understanding, of both methods. The emergence of mixed methods as a gold standard for evaluation will require those already trained in one method to engage with researchers who use other methods. Recognizing this, my advice is to not be afraid of mixed methods. Rather, you should seek ways to develop understanding of new methodologies to enhance your skills and abilities. Doing so will make you a better evaluator and researcher.

Learning Activities

1. Which of the points of integration (complimentary, developmental, initiatory, expanding, triangulating, and mixing) are each of the following examples:
 - a. A large-scale survey of adolescents with epilepsy provided information on adherence levels and associations whilst interviews and focus groups allowed us to engage with individual experiences of chronic illness and medication in adolescence.
 - b. A focus group conducted with a group of adolescents with epilepsy identified mobile phone technology as a potentially important tool in adherence support. We then developed a mobile phone 'app' that reminds patients to take their medication and conducted an intervention study to assess its impact on adherence levels.
 - c. We conducted a clinical study measuring drug levels in individuals and documented self-reported adherence. Video diaries were used to confirm adherence levels.
 - d. A review of case notes found adherence levels of over 90% in a clinic population; however, semi-structured interviews with peer researchers revealed lower levels of adherence and barriers to open discussion with clinicians. We investigated sub-population traits among the non-adherence to identify barriers to adherence.
 - e. A survey of adolescents with epilepsy demonstrates poor levels of adherence. Semi-structured interviews with a sub-group of those surveyed allowed us to explore barriers to adherence.

Additional Readings

- Tariq & Woodman. (2013). "[Using mixed methods in health research.](#)"
- Schoonenboom & Johnson. (2017). "[How to Construct a Mixed Methods Research Design.](#)"
- McKim. (2015). "[The Value of Mixed Methods Research: A Mixed Methods Study.](#)"
- Sockolow et al. (2016). "[Using Mixed Methods in Health Information Technology Evaluation.](#)"
- Dopp et al. (2019). "[Mixed-method approaches to strengthen economic evaluations in implementation research.](#)"
- Cochrane & Davey (2017). "[Mixed-Methods Evaluation of a Healthy Exercise, Eating, and Lifestyle Program for Primary Schools.](#)"

RECRUITING PARTICIPANTS

Learning Objectives

- Identify methods to recruit participants into your study.

Part 1

How do you recruit participants?



Source: <https://www.analyticsindiamag.com/5-key-personality-traits-every-analytics-leader-should-have/>

Probability Sampling

In epidemiological research there is a lot of emphasis on generalizability and random sampling. In health service research there is a recognition that random sampling is not always possible, nor is it always preferred. In this chapter, I want to review the various sampling strategies available to you as a health service researcher and evaluator. To begin, you should recall that in traditional epidemiological studies, a sample statistic is used to estimate a population parameter. Population parameters are the true characteristics that you find in the population, while statistics are the characteristics of your smaller sample. In epidemiology, the goal is to use statistics to describe populations: you are primarily concerned with:

- **Accuracy** refers to the closeness of a measured value to a standard or known value. For example, if in lab you obtain a weight measurement of 3.2 kg for a given substance, but the actual or known weight is 10 kg, then your measurement is not accurate. In this case, your measurement is not close to the known value.
- **Precision** refers to the closeness of two or more measurements to each other. Using the example above, if you weigh a given substance five times,

and get 3.2 kg each time, then your measurement is very precise. Precision is independent of accuracy. You can be very precise but inaccurate, as described above. You can also be accurate but imprecise.

- A **representative** sample is one that has strong external validity in relationship to the target population the sample is meant to represent. As such, the findings from the survey can be generalized with confidence to the population of interest.

To achieve these characteristics, you can use probability sampling methods, which require each person in a population to have a known (non-zero) chance of being chosen for the sample. Probability sampling methods include the following:

- **Simple random sampling** refers to any sampling method in which all individuals in a population have an equal probability of being selected into the study and all possible combinations of individuals in a population is possible. There are many ways to obtain a simple random sample. One way would be the lottery method in which each individual is assigned a unique number and then individual numbers are randomly selected. This type of sampling is good for when you want to report on the incidence or prevalence of an outcome or when you want the selection of your participants to be unbiased.
- **Stratified sampling** begins with dividing the population into groups, base on some characteristic of interest and then selecting individuals randomly from within each stratum. For example, as an example, suppose we conduct a national survey. We might divide the population into groups or strata, based on geography - north, east, south, and west. Then, within each stratum, we might randomly select survey respondents.
- **Cluster sampling** is similar in that individuals are assigned to groups (maybe a school within a school district) and then the groups (also called clusters) are randomly selected.
- **Multistage sampling** involves mixing two or more probability sampling methods together. For instance, you might first begin with randomly selecting clusters and then use simple random sampling to identify individuals within those clusters who you want to survey.

- Finally, **systematic random sampling** uses a list of members in a population and then selects every k^{th} person – where k is a randomly selected number.

In each of these cases, individuals have an equal probability of being selected. However, there are two very important limitations to recognize when you are treating a sample as if it were a probability sample: non-response and non-coverage.

- **Non-response** arises when people who have been selected into the study choose not to participate. As soon as a participant declines to participate, it is possible that your results are biased – because some factor associated with declining to participate might shape your outcome. For instance, wealthier individuals might have less incentive to participate in a study, younger individuals might be less likely to want to complete a survey over the phone and so on and so forth.
- **Non-coverage** is similar to non-response but arises from the fact that not all persons are included in your sampling frame. Even in a population census, homeless individuals might be missed and therefore not included as eligible participants if the census were used as the list from which individuals are sampled. Recognizing these limitations, it should be clear that no survey is every truly perfect – no matter what authors might try to say about their sampling methodology.

Non-probability Sampling

In evaluation research, the goal is to make a judgement about a program. The unit of evaluation is therefore a program, and while individuals may help you understand how to improve a given program – you are rarely trying to estimate population parameters with statistics. Indeed, sometimes it doesn't really matter if your sample is not probabilistically created. In health research and evaluation, you might actually prefer a non-random sample. Indeed, for that reason there are many established sampling methodologies that can be relied upon to help you conduct a study. While it is difficult to draw conclusions about population parameters from these methods – often time your research question will allow you the flexibility to ignore whether or not your results are accurate, precise, or representative. Under these circumstances, the following non-probability samples are available to you:

- **Quota sampling** uses quotas – often proportional to the population being studied – and recruit’s participants to fill these quotas. Quota sampling is particularly useful when you are unable to obtain a probability sample, but you are still trying to create a sample that is as representative as possible of the population being studied. In this respect, it is the non-probability based equivalent of the stratified random sample.
- **Convenience sampling** involves recruiting convenient participants. For example, you might ask participants waiting in your lobby to complete a survey simply because it is easy to get in touch with them. Convenience sampling is very easy to carry out with few rules governing how the sample should be collected. Furthermore, the relative cost and time required to carry out a convenience sample are small in comparison to probability sampling techniques. This enables you to achieve the sample size you want in a relatively fast and inexpensive way. The convenience sample may help you gathering useful data and information that would not have been possible using probability sampling techniques, which require more formal access to lists of populations.
- **Self-selection sampling** is a sampling strategy in which participants opt in to a study – perhaps because they see a flier or social media post. Since the potential research subjects (or organisations) contact you this can reduce the amount of time necessary to search for appropriate units (or cases); that is, those individuals or organisations that meet the selection criteria needed for your sample. The potential units or cases (individuals or organisations) are likely to be committed to take part in the study, which can help in improving attendance (where necessary), and greater willingness to provide more insight into the phenomenon being studied (e.g., a respondent may be more willing to spend the time filling in qualitative, open-ended questions in an online survey, where others may leave them blank).
- **Snowball sampling** is a form of chain referral sampling that involves invitations being provided through social networks and word of mouth. Snowball sampling might be useful if you are hoping to understand something about social networks or communities as a way to understand intergroup or intra group dynamics that might impact your intervention. Similarly, some populations that we are interested in studying can be hard-

to-reach and/or hidden. These include populations such as drug addicts, homeless people, individuals with AIDS/HIV, prostitutes, and so forth. Such populations can be hard-to-reach and/or hidden because they exhibit some kind of social stigma, illicit or illegal behaviours, or other trait that makes them atypical and/or socially marginalized. Snowball sampling is a non-probability based sampling technique that can be used to gain access to such populations because it leverages the social capital of participants to recruit other participants.

- **Maximum variation sampling**, also known as heterogeneous sampling, is a purposive sampling technique used to capture a wide range of perspectives relating to the thing that you are interested in studying; that is, maximum variation sampling is a search for variation in perspectives, ranging from those conditions that are view to be typical through to those that are more extreme in nature. By conditions, we mean the units (i.e., people, cases/organizations, events, pieces of data) that are of interest to the researcher. These units may exhibit a wide range of attributes, behaviors, experiences, incidents, qualities, situations, and so forth. The basic principle behind maximum variation sampling is to gain greater insights into a phenomenon by looking at it from all angles. This can often help the researcher to identify common themes that are evident across the sample.
- **Homogeneous sampling** is a purposive sampling technique that aims to achieve a homogeneous sample; that is, a sample whose units (e.g., people, cases, etc.) share the same (or very similar) characteristics or traits (e.g., a group of people that are similar in terms of age, gender, background, occupation, etc.). In this respect, homogeneous sampling is the opposite of maximum variation sampling. A homogeneous sample is often chosen when the research question that is being address is specific to the characteristics of the particular group of interest, which is subsequently examined in detail.
- **Typical case sampling** is a purposive sampling technique used when you are interested in the normality/typicality of the units (e.g., people, cases, events, settings/contexts, places/sites) you are interested, because they are normal/typical. The word typical does not mean that the sample is representative in the sense of probability sampling (i.e., that the sample shares the same/similar characteristics of the population being studied).

Rather, the word typical means that the researcher has the ability to compare the findings from a study using typical case sampling with other similar samples (i.e., comparing samples, not generalizing a sample to a population).

- **Extreme (or deviant) case sampling** is a type of purposive sampling that is used to focus on cases that are special or unusual, typically in the sense that the cases highlight notable outcomes, failures or successes. These extreme (or deviant) cases are useful because they often provide significant insight into a particular phenomenon, which can act as lessons (or cases of best practice) that guide future research and practice. In some cases, extreme (or deviant) case sampling is thought to reflect the purest form of insight into the phenomenon being studied.
- **Critical case sampling** involves selecting a small number of important cases - cases that are likely to "yield the most information and have the greatest impact on the development of knowledge" Critical case sampling is a type of purposive sampling technique that is particularly useful in exploratory qualitative research, research with limited resources, as well as research where a single case (or small number of cases) can be decisive in explaining the phenomenon of interest. It is this decisive aspect of critical case sampling that is arguably the most important. To know if a case is decisive, think about the following statements: If it happens there, it will happen anywhere; if it doesn't happen there, it won't happen anywhere; If that group is having problems, then we can be sure all the groups are having problems? Whilst such critical cases should not be used to make statistical generalizations, it can be argued that they can help in making logical generalizations. However, such logical generalizations should be made carefully.
- **Total population sampling** is a type of purposive sampling technique where you choose to examine the entire population (i.e., the total population) that have a particular set of characteristics (e.g., specific experience, knowledge, skills, exposure to an event, etc.). In such cases, the entire population is often chosen because the size of the population that has the particular set of characteristics that you are interest in is very small. Therefore, if only a small number of units were included in the sample, it may be felt that a significant piece of the puzzle was missing.

- **Expert sampling** is where you draw your sample from experts – or key informants – in the field you’re studying. It’s used when you need the opinions or assessment of people with a high degree of knowledge about the study area. When used in this way, expert sampling is a simple sub-type of purposive sampling. “Expert” doesn’t necessarily have to mean highly education and skilled in a field. Individuals with lived experience are often the best and most informed *experts* on a variety of issues.

As has already been partially review, there are a variety of advantages of purposive sampling. For instance, one of the major benefits of purposive sampling is the wide range of sampling techniques that can be used across such qualitative research designs; purposive sampling techniques that range from homogeneous sampling through to critical case sampling, expert sampling, and more. Whilst the various purposive sampling techniques each have different goals, they can provide researchers with the justification to make generalizations from the sample that is being studied, whether such generalizations are theoretical, analytic and/or logical in nature. However, since each of these types of purposive sampling differs in terms of the nature and ability to make generalizations, you should always take care in ensuring that you understand how the selection of your participants might influence the validity of your study.

Inviting Participants to Participate

In addition to designing your recruitment strategy, you also need to figure out how you will invite participants to participate. Invitations to participate in your study provide a “first impression” of who you are and what you’re trying to do. Saying the wrong thing can thus drive people away. Remember that people like a personal touch (e.g., Dear Kiffer...) and they like to feel special (e.g., “You have been selected!” or “We’d love to hear from you!”), but they are also busy so make sure you provide the key information about what their participation will mean, when the deadline will be, what you’re aiming to show with your results, and what they will get out of the survey. Most ethics review boards will also want you to disclose any potential harms that someone might encounter during a study. How you invite participants will largely depend on how you will be collecting data from them. We will discuss the various data collection methods later in this text.

Learning Activities

1. An auto analyst is conducting a satisfaction survey, sampling from a list of 10,000 new car buyers. The list includes 2,500 Ford buyers, 2,500 GM buyers, 2,500 Honda buyers, and 2,500 Toyota buyers. The analyst selects a sample of 400 car buyers, by randomly sampling 100 buyers of each brand. Is this an example of a simple random sample?
 - a. Yes, because each buyer in the sample was randomly sampled.
 - b. Yes, because each buyer in the sample had an equal chance of being sampled.
 - c. Yes, because car buyers of every brand were equally represented in the sample.
 - d. No, because every possible 400-buyer sample did not have an equal chance of being chosen.
 - e. No, because the population consisted of purchasers of four different brands of car
2. Suppose you are a hospital administrator and you want to investigate wait times for your emergency room. You decide to sample 1,000 patient records from the 13,000 admissions that occurred over the past year. You ask your records officer to generate a list of 1,000 random numbers between 1 and 13,000. For each, random number generated, you pull the record that was numbered as such on your master list, which was organized alphabetically. What kind of sampling procedure was this?
3. A _____ sample may start with randomly selecting clusters within a population and then selecting individuals randomly from each strata.
4. True or False: Non-response is not a problem if participants are randomly selected in the first place.
5. Identify the type of sampling used in each example below.
 - a. Every 7th person who enters the emergency room is surveyed.
 - b. A researcher randomly interviews 50 people living with HIV and 50 people without HIV.
 - c. All doctors from 30 randomly selected Canadian hospitals were interviewed
6. Dr. Ivar Asbjørn Følling was a Norwegian physician and biochemist who was studying a hereditary metabolic disorder known today as Phenylketonuria (PKU). PKU is associated with intellectual disability as well as other physical problems. The first cases of PKU were identified when a mother had two children with severe disability. She believed that the musty odor associated with her infants was a sign of an underlying cause of the disease afflicting her children. Dr. Følling conducted ferric chloride urine tests and saw that their urine turned green -- which he never had seen before. As he hoped to identify more cases with PKU he began walking through hospital wards, asylums, and other institutions and used his nose to identify potential subjects -- looking for that familiar musty smell. In doing so, Dr Følling used _____ sampling to identify those who would help him to study the disease he had discovered.

Additional Readings

- Palinkas et al. "Purposeful sampling for qualitative data collection and analysis in mixed methods research."

Key Resources & Guides

- W. K. Kellogg Foundation. (2017). "The Step-by-Step Guide to Evaluation: How to Become Savvy Evaluation Consumers." W. K. Kellogg Foundation.
- Harris (2017). "Evaluating Public and Community Health Programs." Jossey-Bass.
- Public Health Ontario. (2016). "Evaluating health promotion programs: introductory workbook." Public Health Ontario.
- World Health Organization. (2013). "WHO Evaluation Practice Handbook." WHO.
- Watkins et al. (2012). "A Guide to Assessing Needs Essential Tools for Collecting Information, Making Decisions, and Achieving Development Results." The World Bank.
- Bowen et al. (2012). "A Guide to Evaluation in Health Research." Canadian Institutes of Health Research.
- European Centre for Disease Prevention and Control (2011). "Technical Report: Evidence-based methodologies for public health." European Centre for Disease Prevention and Control.
- 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." Centers for Disease Control and Prevention.
- World Health Organization. (2007). "Drinking and Driving: A Road Safety Manual. Chapter 4: How to Evaluate the Programme." World Health Organization.
- Bertrand. (2006). "Fundamentals of Program Evaluation." JHSPH Open.
- W. K. Kellogg Foundation. (2004). "Logic Model Development Guide." W. K. Kellogg Foundation.
- U.S. Department of Health and Human Services Centers for Disease Control and Prevention. (1999). "Framework for Program Evaluation in Public Health." MMWR.

Appendix 1: Informal Participatory Data Collection Strategies

Contents

Appendix 1, Part I: Considerations/Questions to Decide Which Methods to Use	176
Appendix 1, Part II: Consultation Methods: Methods for Idea Generation & Prioritization	177
Card Collection/ Clustering	177
Dotmocracy	177
Nominal Group Technique	177
OPERA	178
Structured Brainstorming/ Structured Ideation	178
Think-Pair- Share	179
Snowball	179
Appendix 1, Part III: Methods for Exploring Issues	180
Anecdote Circles	180
Brain Writing	180
Fish Bowl	180
Focus Groups	181
Mindmapping	181
Round Robin	181
Visual Notetaking	182
World Café	182
Appendix 1, Part IV: Methods for Feedback and Assessing Participants’ Interest, Experience, or Opinions	183
Democracy Wall	183
Human Spectrogram	183
Spider Diagram	184
Appendix 1, Part V: Broad Models and Techniques for Consultation and Engagement	185
Appreciative Inquiry	185
Community Visioning	185
Open Space	186

Appendix 1, Part I: Considerations/Questions to Decide Which Methods to Use

Purpose/objectives	<ul style="list-style-type: none"> • Desired outcomes
Topic	<ul style="list-style-type: none"> • What is the scope of the topic? • How specialized or esoteric is the topic? • What level of knowledge is required about the topic? • Is the topic controversial or contentious? • What is the relevance of the topic to the participants?
Setting/location	<ul style="list-style-type: none"> • Space • Comfort level <ul style="list-style-type: none"> ○ Emotional ○ Physical (e.g., heating) ○ Distractions • Are participants traveling far?
Participants/audience	<ul style="list-style-type: none"> • Size • Participant characteristics <ul style="list-style-type: none"> ○ Demographics (including gender, age, disability, literacy, spoken/understood languages) ○ Diversity within the group ○ Culture ○ Role/level ○ Knowledge/expertise ○ Motivation to participate/incentives ○ Political considerations/history of working with the participants • Group dynamics <ul style="list-style-type: none"> ○ How well do the participants know each other? ○ How have the participants worked together in the past? ○ What is the level of trust within the group? <ul style="list-style-type: none"> ▪ At what stage in the project/experience working together is the consultation taking place?
Resources	<ul style="list-style-type: none"> • Available materials • Access to technology • Staff/helpers for the session • Facilitators <ul style="list-style-type: none"> ○ Number ○ Experience with methods ○ Experience with the group
Time	<ul style="list-style-type: none"> • How long do you have with the group? • What else is going on that day? • What time of day? (e.g. first thing in the morning, after lunch, etc.)

Appendix 1, Part II: Consultation Methods: Methods for Idea Generation & Prioritization

Method	Description	Works well for...	Other considerations
Card Collection/ Clustering	All participants are given a stack of cards or post-it notes on which they can write their ideas. Cards are collected by a facilitator and ideas are read out loud. The group works to put the ideas on a wall in categories – adjusting the categories or clusters of ideas as they go.	<ul style="list-style-type: none"> Getting a lot of ideas heard quickly without discussion being dominated by specific individuals Getting a sense of the full scope of ideas held by a group 	<ul style="list-style-type: none"> Other methods can be used to prioritize these ideas or the clusters afterwards.
Dotmocracy	A common method to prioritize ideas or topics by a large group of people. Participants are given a certain number of dot stickers that they can place next to ideas or topics written on sheets affixed to the walls of a meeting room. Ideas receiving the most numbers of dots are the ones that are then taken up by the group for future work or activities.	<ul style="list-style-type: none"> A large group Moving forward quickly when a lot of ideas are on the table Getting participation from all individuals (not just the loudest voices) 	<ul style="list-style-type: none"> Can change numbers of dots, instructions for placing them, etc. Can be used in conjunction with many idea generation methods
Nominal Group Technique	Nominal Group Technique can help a group to generate ideas or suggestions for a specific problem and make a decision about how to move forward. In this method, each participant proposes their idea(s). The facilitator helps the group to eliminate duplicate ideas and respond to any questions about the ideas. Ideas are then ranked by all participants independently (in one or two rounds), with the highest ranked idea taken as the group's decision.	<ul style="list-style-type: none"> Involving all participants (regardless of personal characteristics or group dynamics) Controversial topics Generating a large number of ideas 	<ul style="list-style-type: none"> Can be used in groups of all sizes. Can have participants contribute one or multiple ideas. Ranking/voting can be done in many ways

Method	Description	Works well for...	Other considerations
OPERA (Own, Pair, Explain, Rank, Arrange)	Method to elicit, collaboratively develop, and prioritize ideas. Participants first write their own ideas down independently (own). Then they are grouped in pairs to discuss their ideas (pair). Pairs produce a few ideas (up to 4) that they agree with, write them on a flipchart or post-its, and briefly describe them to the large group (explain). The pairs then come back together to discuss all of the ideas that were presented and decide which ones they like the most (only allowing them to select one from their own flipchart) – they can select up to 25% of the presented ideas (rank). All pairs are given a different colour marker or sticker and mark the flipcharts with their selected ideas. The facilitator tallies the votes and then leads a discussion with the large group to categorize the highest ranked ideas by category or theme (arrange).	<ul style="list-style-type: none"> • Involving all participants (regardless of personal characteristics or group dynamics) • Keep participants engaged/active in all steps of the process • Getting a lot of ideas on the table and not narrowing the range of ideas too quickly 	<ul style="list-style-type: none"> • Make sure you allow for enough time for all the steps in the process (about 1 hour) • Not for use with really large groups (too many pairs)
Structured Brainstorming/ Structured Ideation	<p>The first step involves <i>idea generation</i> to get as many ideas on the table as possible. Going around a circle or room, each participant contributes one idea that the facilitator writes up on a flipchart. Other group members are not permitted to critique, build on, or question the ideas. The facilitator repeats the round until participants do not have any other ideas to contribute.</p> <p>The next step is <i>idea evaluation</i>. The facilitator helps the group to eliminate duplicates or ideas that are out of scope, group similar ideas, and then sort ideas so that they can be worked with in future planning activities.</p>	<ul style="list-style-type: none"> • Encouraging all participants to participate in non-judgemental environment • Eliciting a wide range of ideas; avoiding “group think”; reducing reluctance to share incomplete or “risky” ideas 	<ul style="list-style-type: none"> • The larger the group, the longer this activity will take • One way to categorize ideas is to divide them into ideas of immediate usefulness, areas for further exploration, and radically new approaches to the problem. • Can also use another prioritization/selection process (e.g., dotmocracy)

Method	Description	Works well for...	Other considerations
Think-Pair-Share	Participants first work independently to write down their response to a question or ideas on a specific topic. Then they break up into pairs to discuss their ideas or solutions. One member of each pair reports back to the full group about the ideas and the pair's discussion.	<ul style="list-style-type: none"> Involving all participants (regardless of personal characteristics or group dynamics) Bringing up a wide range of ideas 	<ul style="list-style-type: none"> For really large groups, may want to adapt plans for the share activity
Snowball	Present an idea, question, or issue to participants. Each participant first thinks about the idea/question/issue for one minute, with the goal of generating at least three reactions, comments, answers, etc. Two participants then come together with their lists and try to come up with three things they agree on. The pairs of participants then join with another pair, and try to come up with three things they agree on. Repeat for as many iterations as desired. Eventually, bring the class together as a group to hear what the participants have decided are the three most important issues, questions, ideas relevant to the topic discussed.	<ul style="list-style-type: none"> Involving all participants (regardless of personal characteristics or group dynamics) Bringing up a wide range of ideas Coming to consensus quickly with a large group 	<ul style="list-style-type: none"> Gets unwieldy after a certain point (if group is too big)

Appendix 1, Part III: Methods for Exploring Issues

Method	Description	Works well for...	Other considerations
Anecdote Circles	A small group method to explore issues and ideas. Similar to focus groups except participants are asked to share stories from their own experiences, instead of stating their opinions or proposing solutions.	<ul style="list-style-type: none"> • Eliciting participants' direct experiences and examples • Smaller groups (6-8 is ideal) 	<ul style="list-style-type: none"> • Requires a strong facilitator
Brain Writing	Each participant writes down an issue or question that needs answering and then passes their papers to the participant on the left. That participant writes a suggestion or answer below and passes to the next participant on his/her left. The next participant can either add to the first suggestion or write down a new suggestion or answer. The question goes around the circle this way until it returns to the original question/issue writer. An optional debrief round can follow where the participants share their problems and the best solutions they received.	<ul style="list-style-type: none"> • Answering multiple questions or developing solutions to multiple issues • Eliciting ideas from all group members (introverts and extroverts) 	<ul style="list-style-type: none"> • Set up: room for small groups (6-8 people) sitting in a circle. • Important to frame the question well
Fish Bowl	Participants sit on chairs set up "in the round," with an inner circle of 3-5 chairs. The experts or conversation leads sit in the chairs in the middle and the facilitator poses a question or topic for them to discuss. The facilitator can take questions or comments from the audience for the people in the fishbowl to respond to.	<ul style="list-style-type: none"> • Allows meeting participants to hear and contribute to conversation and discussion amongst a small group of experts. • Discussing contentious or controversial topics • Up to 50 participants when there are some "experts" or leaders who have specific information or perspectives to share • Avoiding lengthy presentations 	<ul style="list-style-type: none"> • Take a few minutes in the beginning to explain the method and the topic to be discussed • Many variations for this method exist to get participants involved in asking and answering questions

Method	Description	Works well for...	Other considerations
Focus Groups	A method to collect information from a small group about the attitudes, perspectives, or knowledge about an issue or topic. Often used in research or evaluation contexts, but can also be used for consultations.	<ul style="list-style-type: none"> • Understanding a complex issue that benefits from participant interaction to elicit additional insights and information • Small groups (5-10 people) 	<ul style="list-style-type: none"> • Requires a skilled facilitator • Personality characteristics and group dynamics may affect the degree of participation by all participants • Contentious issues may be difficult to explore in a group
Mindmapping	A note-taking method that allows participants to collaboratively create a visual map to display components and relationships between different aspects of an issue or topic.	<ul style="list-style-type: none"> • Engaging participants who are visual learners/thinkers • Quickly capturing wide-ranging ideas • Groups for which there may be language or literacy barriers 	<ul style="list-style-type: none"> • Can be hand-drawn on flipcharts or generated using specific computer software • End product can be shared/owned by all participants
Round Robin	The session is broken up into separate segments and several tables are set up, each designated for a specific component or aspect of a single topic. Small groups of participants rotate between tables during the different sessions, brainstorming and recording their ideas. In each segment, the small group at the table considers and adds to ideas/conversations that have taken place in the segments before them. In the last segment, the groups go back to their original tables to review all the ideas. Then each group reports back to the full group with a summary of all the discussions and ideas.	<ul style="list-style-type: none"> • A medium-sized group to allow for 4-6 people in each small group and one group at each table • Allowing participants to learn and talk about a range of issues or angles on a topic in small group format • Advancing discussions or planning in more depth than occurs in single small or large group conversations 	<ul style="list-style-type: none"> • Quite similar to a World Café but the tables do not have a designated host or facilitator • To keep the session fun and engaging, can reduce the time allotted for each successive session

Method	Description	Works well for...	Other considerations
Talk Show	A facilitator acts as a talk show host with 3-4 guests sitting in the front of the room. The host introduces the topics and poses relevant questions to the guests, who respond. The other participants serve as the audience, who are invited to ask questions of the guests or share their own perspectives. At the end of the session, the host provides a summary reflection of the conversation.	<ul style="list-style-type: none"> • Exploring issues where there are a small group of participants or experts with more knowledge • Sharing knowledge in a fun and engaging format • Medium-sized groups, to permit for questions/comments by a large proportion of participants 	<ul style="list-style-type: none"> • Can be carried out formally or quite tongue in cheek • Talk show guests should have different areas of knowledge or perspectives/opinions on the topic • Host should be engaging and energetic
Visual Notetaking	A designated person takes notes throughout a meeting, representing the discussion and decisions with a mix of drawings and words to create a comprehensive visual representation of the meeting.	<ul style="list-style-type: none"> • Creating a visually engaging product from a group consultation or planning process • Representing different aspects or components of a topic or project 	<ul style="list-style-type: none"> • Requires a skilled visual note taker • Result can be used to introduce or promote a project or organization to stakeholders or new team members
World Café	The session is broken up into separate segments and several tables are set up, each with a facilitator prepared to lead a discussion about a specific issue or topic assigned to the table. Participants rotate between tables during the different sessions, building on the conversations that have taken place in the segments before them. At the end of the session, the facilitators report back to the full group about what emerged in the discussions through all the segments.	<ul style="list-style-type: none"> • Exploring an issue or related issues in some depth • A medium-sized group to allow for 4-5 people plus one facilitator at each table • Allowing participants to learn and talk about a range of issues or angles on a topic in small group formats • Advancing discussions or planning in more depth than occurs in single small or large group conversations 	<ul style="list-style-type: none"> • Requires about 60-90 minutes

Appendix 1, Part IV: Methods for Feedback and Assessing Participants' Interest, Experience, or Opinions

Method	Description	Works well for...	Other considerations
Democracy Wall	<p>Flip chart sheets are placed on a wall of the meeting room with various prompts and participants can use markers or post-it notes to add their responses. A common array of prompts includes the following statements:</p> <ul style="list-style-type: none"> • I discovered that ... • I noticed that... • I felt that... • I learned that ... • I would like to suggest... <p>You can also post sheets with other prompts, such as:</p> <ul style="list-style-type: none"> • I'm still confused about... • I am frustrated by... • I'd like to learn more about... • We didn't have enough time to discuss... • I really liked/appreciated... 	<ul style="list-style-type: none"> • Encouraging feedback from all participants over the course of a meeting or event • Creating a shared pool of reflections • Affirming participants' experiences (through seeing others' posts and allowing meeting organizers/facilitators to respond to them during the meeting) 	<ul style="list-style-type: none"> • Think about how to ensure some confidentiality for participants who post (e.g., everybody uses the same colour post-it or marker or gets to choose from a basket placed near the wall; there is unstructured time when participants can look at and contribute to the wall)
Human Spectrogram	<p>A method that demonstrates degree of interest or expertise of a group of participants about a specific idea or approach. The facilitator denotes a line on the floor of the room that represents a Likert scale with specific points as the anchors. Participants are asked to line up at the anchor that corresponds to their opinion or response to create a human bar graph.</p>	<ul style="list-style-type: none"> • Getting everybody involved • A quick, active, and fun way to get a sense of opinions in the room • Getting feedback or assessing expertise or interest on a topic 	<ul style="list-style-type: none"> • Participants' responses are not confidential • Can do spectrograms for several questions in quick succession • May want to have a way to record or draw the results of the spectrogram

Method	Description	Works well for...	Other considerations
Spider Diagram	<p>A method to quickly get feedback on several aspects of a workshop or meeting progress. The facilitator draws a circle on a flipchart. Around the outside of the circle, write the aspect or component you would like to evaluate and draw a line from the centre of the circle to the label (this will create several radii for the circle). Connect the radii by drawing lines between them at set distances from the centre (your figure will now look like a bit like a spider web). Participants then rate progress or provide their feedback for each of the aspects or components, by marking or placing a dot at the place on the radius line that corresponds to their rating (with highest ratings towards the centre and lowest ratings at the outside of the circle). The facilitator then leads a discussion of the results with the group.</p>	<ul style="list-style-type: none"> • Getting and discussing feedback in a quick and interactive format • Visually demonstrating the range of opinions in a group • Assessing the progress of a meeting or project 	<ul style="list-style-type: none"> • Can be used as an active method by creating the spider web on the floor with masking tape and asking participants to stand at the points corresponding to their ratings of various aspects • Can create spider diagrams at various points during a meeting or project to mark progress or compare opinions over time

Appendix 1, Part V: Broad Models and Techniques for Consultation and Engagement

Method	Description	Works well for...	Other considerations
Appreciative Inquiry	<p>An approach to project planning and team building that builds on existing strengths of people, projects, and organizations. More than a facilitation method, it is a five-step process that involves:</p> <ul style="list-style-type: none"> • Definition (establish the focus/scope of the inquiry) • Discovery (elicit positive stories about the system – first in pairs and then shared with larger groups) • Dream (collective visioning for the future) • Design (build on the positive aspects of the past and present state and plan to create the desired future) • Destiny (put an action plan in place to make it happen). 	<ul style="list-style-type: none"> • Contentious or controversial topics • Supporting change in an organization or community • Building constructive and supportive relationships within teams 	<ul style="list-style-type: none"> • Also used as a model or framework in research and evaluation
Community Visioning	<p>An approach often used for strategic and urban planning initiatives that engages participants to imagine and explore changes and future scenarios, along with thinking through the steps and supports needed to achieve this vision.</p> <p>Visioning processes engage participants through independent reflection, sharing in small and large groups, and distilling common ideas about obstacles, opportunities, and pathways to get to the future vision through a variety of active learning and engagement methods.</p>	<ul style="list-style-type: none"> • Large and/or diverse groups of participants • Building a sense of community or collaboration • Focusing on strengths and moving forward from conflict or controversy • Generating a large number of ideas 	<ul style="list-style-type: none"> • For large groups, you need a number of facilitators to capture participants' ideas • Requires a plan for follow-up so that participants feel engaged in next steps and that their ideas will be useful for some future work or process

Method	Description	Works well for...	Other considerations
Open Space	<p>An approach to collective problem solving or planning in which participants engage in simultaneous mini-discussions about a theme or topic based on questions or issues identified by group members at the start of the meeting. Open space meetings do not have a pre-set agenda; instead participants self-select into small discussion groups and move from group to group whenever they decide they would like to learn or contribute elsewhere.</p> <p>Open Space incorporates the four guidelines or principles to seek to maximize and recognize creative and collaborative processes:</p> <ol style="list-style-type: none"> 1. Whoever comes are the right people 2. Whatever happens is the only thing that could have 3. Whenever it starts is the right time 4. When it's over, it's over 	<ul style="list-style-type: none"> • A wide range of group sizes (from 5 to over 1000) • Complex situations, or when conflict or controversy is impeding project planning or progress • Building energy, motivation, and shared leadership/ownership • Helping groups to make decisions quickly • Participants who already have a vested interest or motivation for advancing a project or process 	<ul style="list-style-type: none"> • Time required can range from a couple of hours to a few days; practitioners say that the longer the space is open, the more transformative it becomes • Group members take responsibility for facilitating discussion on specific areas/issues of interest and communicating about the plans or ideas that emerge from the discussion • Open space is “an emergent process”: it may be chaotic or disorganized at times, but order will return eventually if the group is allowed to take its time to go through the process

Key Resources for Appendix 1

- [Better Evaluation](#)
- [Engaging Queenslanders: A guide to community engagement methods and techniques](#)
- [Imagine Chicago Tools for Engagement](#)
- [Institute for Development Studies: Participatory methods website](#)
- [Knowledge Bucket Wiki](#)
- [KS Toolkit Wiki](#)
- [Participatory Methods Toolkit: A practitioner's manual](#)
- [Tools and techniques for EfS and Stakeholder Engagement programs](#)
- [Examples of Active Learning Strategies](#)

Towards the Public's Health: Methods in Health Research and Evaluation was prepared by Dr. Kiffer G. Card as the primary text for Health Research and Evaluation Courses. This First Edition has not been extensively edited or referenced. The activities have not been reviewed and are still under development. A Second Edition of this text is expected in 2020 to correct these shortcomings. If you have feedback about this text or are interested in collaborating as an author, editor, or publisher, please contact Dr. Card at kiffcard@gmail.com.