1 EVALUATION MATTERS

What we know is a drop, what we don't know is an ocean.

—**Isaac Newton**

LEARNING OBJECTIVES

1.1 Define evaluation and identify programs and policies that might be evaluated.

1.2 Describe the purpose of evaluation and its relationship to research.

1.3 Identify the guiding principles, standards, and competencies that govern the field of evaluation.

1.4 Distinguish between formative and summative evaluation.

1.5 Compare and contrast internal and external evaluation.

1.6 Explain the embedded evaluation model.

1.1 WHAT IS EVALUATION?

Welcome to the field of evaluation! Whether you are new to the field and this is your first course in evaluation or you are a seasoned evaluator looking to explore new approaches, we are thrilled to walk with you on this journey and hope that you find the material in this text helpful to you. So a good place to start . . . what is evaluation? Merriam-Webster online defines evaluation as the "determination of the value, nature, character, or quality of something." We all do that on a daily basis:

- We estimate if a product is worth buying (are the upgraded features on this new iPhone version worth the additional $100?).
- We judge whether spending extra time on a homework assignment is worth a higher grade or if settling for a lower grade will impact our overall course average.

- We rate our professors (yes, professors really do look at these ratings from time to time; I revise my courses every semester based on student feedback).
- We appraise the work ethic and quality of one of our coworkers or fellow students.
- We assess the extent to which we will use a textbook in the future and determine whether it is more advantageous for us to rent the book for the semester or purchase a copy. If a textbook is purchased, we may then assess at the end of course whether it is a book we will keep or one that we will sell.
- We make decisions about whether we can afford to rent an apartment on our own or if we have to get a roommate.

If you have contemplated any of these decisions, you are already an evaluator of sorts. While we will not spend class time debating the merits of the new iPhone, we will provide you with strategies to systematically make evaluative decisions.

I am sure you have noticed the "valu" embedded in the word evaluation. Like many words, "value" has multiple dimensions. The Etymology Dictionary online asserts that the term "value" is derived from the Latin word *valere*, meaning to be well, strong, and of worth. Well and strong relate to merit and have to do with inherent value, while worth is typically interpreted within a certain context. Thus, what is being evaluated, the *evaluand* (Scriven, 1979), may have an inherent value that is free of any context. On the other hand, the evaluand may only be of value to a particular group or in a specific context or at a certain time. In their analysis of the two aspects of value—merit and worth—Lincoln and Guba (1980) use the example of gold. They explain that gold can be judged on its merit or its worth. Judged on its merit, gold has inherent beauty. Judged on its worth, gold has a variable value according to the gold trading markets. Likewise, an SAT or GRE prep course may be judged on its merits based on its coverage of material and clarity of instruction. However, judgments based on worth to you likely relate to how well you performed on the SAT or GRE. While Lincoln and Guba recognize that both the merit and worth of an evaluand can change over time, they emphasize the importance of deliberately considering context, and perhaps even multiple contexts, when making evaluative judgments. So whether you are evaluating a new purchase or a course you are taking, consider both its intrinsic value and its value to you at this time in your life.

We have defined **evaluation** as a method of determining value and the **evaluand** as the subject of the evaluation. While we all evaluate as part of our human nature, this textbook focuses on **program evaluation**. That is, the evaluand is the program and the focus is on determining the merit or worth of that program. For the purposes of this textbook, a **program** is defined broadly to include

- A group of activities;
- Small, focused interventions;
- Organization-wide projects;
- Statewide initiatives;

- National reforms and policies; and
- International programs.

The tools explored in this text are applicable to evaluating a set of activities or small interventions, as well as larger initiatives and multifaceted policies. Examples of programs include focused interventions such as the Olweus Bulling Prevention Program (Olweus & Limber, 2007), schoolwide programs such as Success for All (Slavin et al., 1996), nationwide programs such as Teach for America (Penner, 2021), national reforms such as Head Start (2019), and international programs such as Fulbright (Lally & Islem, 2023). See the "In the Real World" box for several additional examples of programs.

IN THE REAL WORLD . . .

World of Words (WOW) is a curriculum used to supplement the material taught in prekindergarten science lessons. The What Works Clearinghouse (WWC, 2023a) examined four rigorous evaluations of WOW and determined the curriculum to have strong evidence (Tier 1) of positive effects on language development for young children. The WOW curriculum intends to improve language skills through classroom conversations and shared book readings focused on science topics.

Dual Language programs are often used in elementary schools to increase bilingualism, foster appreciation for cultural diversity, and improve academic achievement. These programs include instruction in English and a second language. Typically, 50% of instruction is in English and 50% in the second, or partner, language. The WWC (2022a) examined two evaluations, a randomized controlled trial and a quasi-experimental study, of dual language programs and found moderate evidence (Tier 2) of potentially positive effects on literacy achievement in Grade 5.

The **Good Behavior Game** is a team-based classroom management intervention for grades K–11. The WWC (2023b) examined 16 evaluations of the Good Behavior Game and found strong evidence (Tier 1) of positive effects on student behavior. In addition, the WWC found promising evidence (Tier 3) of positive effects on teacher practice and potentially positive effects on writing conventions and writing productivity at the elementary level.

Growth Mindset interventions aim to increase the postsecondary success of students. Programs to foster a growth mindset are typically implemented with students in their first semester in college; growth mindset programs include education to raise awareness that intellectual ability is not fixed but rather can improve over time. Growth Mindset interventions also employ instruction on strategies to manage challenges students may encounter during their college years. The WWC (2022b) examined six evaluations of Growth Mindset programs and found moderate evidence (Tier 2) of positive effects on academic achievement.

Project QUEST (Quality Employment through Skills and Training) is a postsecondary career and technical education (CTE) intervention. The intervention provides support services to individuals enrolled in occupational training programs to increase credit accumulation, improve program completion rates, and ultimately improve employability and earnings of participants. The WWC (2021) examined three evaluations of Project Quest and found strong evidence (Tier 1) of positive effects on completion of industry-recognized credentials, certificates, or licenses. The WWC also found promising evidence (Tier 3) of potentially positive effects on credit accumulation.

Source: What Works Clearinghouse (http://ies.ed.gov/ncee/wwc)

1.1.1 What Is the Purpose of Evaluation?

Evaluation intends to determine merit and worth. Yet, as discussed in the previous section, evaluation is also contextual. Determining the merit and worth of a program and focusing on its value for a group of people, under a certain set of circumstances, and in a specific context, is no easy task. This is especially true when the consequences of your evaluation might have human and financial implications. Thus, the way we go about evaluating a program is critical to making evaluative determinations. We trust that medical schools effectively evaluate their students to ensure that the cardiothoracic surgeon operating on a loved one has the skill to do so. We often have no choice but to trust that the auto mechanic evaluating why our car broke down is competent at repairing engines and ethical in quoting prices. As a medical school has procedures to evaluate its students and mechanics have protocols to assess car problems, program evaluators have methods, processes, standards, and tools to guide their evaluations.

The evaluative methods I have adopted over the course of my career have been heavily influenced by two factors. The first is my undergraduate training as an engineer. As an engineer, I learned how to think systematically. Engineering is a process and a way of thinking, as well as a discipline. The approach I learned in this context is one that starts with problem identification and description and ends with a set of solutions or recommendations—after which, the process starts again. The second influence on my thinking regarding evaluation is a well-known evaluator named Carol Weiss. Unfortunately, I never met her, but her seminal book titled *Evaluation* (Weiss, 1998) has been a trustworthy companion for decades. Of all who have attempted to define evaluation, Weiss's definition strikes me as the most comprehensive (pp. 4–5):

> Evaluation is a systematic assessment of the operations and/or the outcomes of a program or policy, compared to a set of explicit or implicit standards, as a means of contributing to the improvement of the program or policy.
>
> The primary tenets of Weiss's definition are as follows:

- Evaluation is systematic.
- Evaluation focuses on operations.
- Evaluation focuses on outcomes.
- Evaluation evidence is compared to a standard.
- Evaluation is about improving programs and policies.

Evaluation is a **systematic** examination of a program. It uses the scientific method. Remember that from grade school?! The scientific method has been around since the 1600s, at least, and involves asking a question, researching the question, making and testing a hypothesis, analyzing data, and documenting results. It is what engineers, psychologists, biologists, and social scientists use in their work. It is also the basic science that underlies evaluation. Evaluation is a formal, logical, and organized endeavor undertaken according to a plan.

Evaluation examines the **operations** of a program. The operations of a program include both what is implemented as part of the program and how it is implemented; operations are the processes involved with implementing the activities of a program. Operations are important for two main reasons: interpretation and improvement. Understanding the state of operations allows us to document how a program is operating. Understanding how a program operates, in turn, allows us to determine whether the operations are in fact in accordance with what was intended. This is important for interpreting any results. For instance, if depressive symptoms decrease after a new medication is dispensed, one might be led to believe the medication is effective in treating depression. However, examination of operations might show that over 50% of patients did not take their medication. Without examining how a program operates in actuality, versus how it might have been planned, one can draw inaccurate conclusions. The second reason to examine operations is similar to the first, but involves using the information gathered when looking at operations to make improvements to a program while in operation. So, if through ongoing, systematic evaluation it is determined early in the implementation that patients are not taking their medication, new interventions could be put in place to improve compliance.

Evaluation also examines the **outcomes** of a program. Outcomes are the results that occur during and after implementation of a program. Examining the outcomes of a program allows you to make determinations about the effectiveness of a program. If, for instance, we are examining the overall instructional program at your college or university, a focus on operations might relate to the quality of teaching and the rigor of assignments, but the outcome would likely be student learning. Knowing the extent to which students are learning can help (or hurt) a college with recruitment, affect the success of fundraising campaigns with alumni, and influence partnerships with organizations that may be interested in hiring graduates. By measuring outcomes, we can make determinations about whether the program worked, to what extent, in what ways, and for whom.

Evaluation evidence is compared to a **standard**. A standard is a target or yardstick that informs us of the ideal state. Standards are what we use, implicitly or explicitly, to judge the merit or worth of a program. The standard directs us in making this judgment. In our examples at the start of the chapter, we mentioned purchasing a new cell phone. How much greater would it need to be than the former model for you to purchase it? Likewise, the conflict of how much time to put in on an assignment versus the value of that assignment was introduced. Do you have an implicit standard, one you may not have written down or expressed, that drives whether spending an hour on a 5-point extra credit assignment is worth it to you? People who operate programs and those who evaluate those programs wrestle with similar decisions. If a medication helps 50% of the people who take it, is that enough to continue dispensing the medication? Probably. What if it helps 25%? Or what if it helps 50%, but makes the symptoms for 25% of the people worse? There are no easy answers when it comes to standards, though we explore this thinking further in Chapter 7 in the section on indicators and targets.

Finally, evaluation is not performed in a vacuum and it is not simply an exercise in curiosity. Evaluation is focused on how well a program works, under what conditions, and for which people. Evaluation is intended to provide information aimed at **improving programs and policies**. Ideally, the information obtained through an evaluation will be used to create more effective and efficient programs and policies. And, I imagine for most using this textbook, the programs and policies you might examine are intended to help people. For evaluators, that is the end result. It is the reason we do what we do—to inform programs and policies that will ultimately improve the lives of people. A secondary reason we evaluate programs and policies is to contribute to the field through informing theory and practice.

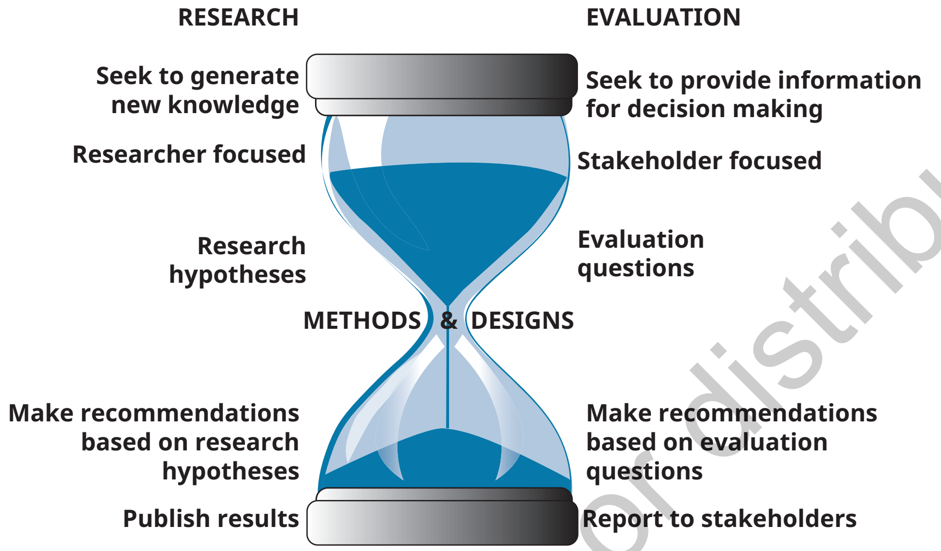
1.1.2 How Is Evaluation Different From Research?

Like evaluation, **research** is a systematic investigation in a field of study. In many ways, evaluation is a form of research. In fact, some refer to evaluation as evaluation research. Evaluation and research use the same methods and designs. The underlying science is the same. However, unlike evaluation, pure research is primarily focused on contributing to the greater body of knowledge in a certain area. This is in contrast to the primary purpose of program evaluation, which is to improve and make decisions about programs and policies.

Evaluation is more practice oriented than research. Evaluation findings are intended for use within a program or policy to effect change. In addition, while researchers often develop their own research hypotheses, evaluators typically work with program staff to develop questions to shape and focus the evaluation. In addition, as stated in the discussion of standards, evaluation intends to compare the evaluation results with what should be. That is, it is judgmental in nature and the eventual intention is to make a decision about whether a program should be continued, expanded, scaled down, or discontinued. Moreover, because evaluators are working in action settings where programs are being implemented in real time, we often face obstacles that might not be encountered in a lab or controlled research setting. For instance, in an evaluation relying on state test scores to examine the impact of curricular changes over a 5-year period, policymakers discontinued use of the state test and instead adopted an assessment that hindered comparisons to previous scores, which in effect, compromised the evaluation. Finally, evaluators depend on people for data collection. As such, interpersonal skills, such as strong communication and listening skills, as well as flexibility and even a positive attitude, can be determinants of whether an evaluation is efficacious or unsuccessful.

However, there are some elements that evaluation and research share. For instance, as stated previously, program evaluation and research use the same methods and designs to frame and conduct their studies. Additionally, like researchers, evaluators have an obligation to disseminate their research. Sometimes this may be publication in peer-reviewed journals, as is common for researchers. However, for both researchers and evaluators, findings should also be shared with individuals or organizations that may benefit from understanding or adopting recommendations, as well as policymakers who are responsible for making policy that may be impacted and improved by the findings. Finally, both evaluators and researchers have ethical obligations and a code of conduct that guide how, why, from whom, and under what conditions data are collected. See Figure 1.1, adapted from a post by Lavelle (2010) on AEA365, a daily blog sponsored

FIGURE 1.1 ■ Research and Evaluation


RESEARCH
EVALUATION
Seek to generate new knowledge
Seek to provide information for decision making
Researcher focused
Stakeholder focused
Research hypotheses
Evaluation questions
METHODS & DESIGNS
Make recommendations based on research hypotheses
Make recommendations based on evaluation questions
Publish results
Report to stakeholders

Source: Adapted from LaVelle, J. (2010, February 26). John LaVelle on describing evaluation [Blog post]. https://aea365.org/blog/john-lavelle-on-describing-evaluation/

by the American Evaluation Association (AEA), for an illustration of some of the differences and intersections between evaluation and research.

While this discussion highlights commonalities and differences between research and evaluation, it should be noted that there is little consensus in the field regarding the interrelationship and intersection between evaluation and research. Wanzer (2020) studied how evaluators and researchers make a distinction between evaluation and research. Findings from their survey of 787 evaluators and researchers revealed that there was no clear common definition that differentiated evaluation from research. Yet about half of respondents said there were differences related to focus, purpose, and context. They felt evaluation focused on different types of questions than research, had a different purpose than research, included a greater consideration of stakeholder needs, and differed from research in context and setting of the study. Some respondents believed that evaluation and research differed by phase of the study, study design, generalization of findings, politics, judgmental nature, or independence. Interestingly, evaluators were more likely to define evaluation and research as intersecting, while researchers were more likely to define evaluation as a subset of research. Wanzer concluded that this lack of consensus presented challenges to the field of evaluation, in terms of communicating the definition of evaluation to others, determining what it means to call oneself an evaluator, and defining (and advancing) evaluation as a discipline. Understanding the intersections and distinctions between research and evaluation will continue to be a topic of discussion for years to come.

1.2 WHY EVALUATE?

A pioneer in the field of research for effective marketing, Arthur C. Nielson based his work on the philosophy that "the price of light is less than the cost of darkness." I know this is deep. Perhaps too deep for the hour of the day (or night) during which you are reading this text. But it is definitely worth the time to think about the implications of his statement. If we are honest with ourselves, it is why we further our education. We go to school, we read, we study because in the long run we believe it will make a difference in our quality of life that is worth the price we pay for our education.

In fact, it would be difficult to find an example where knowledge and truth do not matter. Yet so much *light* is ignored because of the immediate and short-term cost, resulting in a great long-term cost of managing the *darkness*. There is no place this is truer than in policy making. For instance, it is well-documented that drug treatment is a more effective as well as cost-efficient solution than minimum mandatory sentences for and incarceration of drug offenders (McVay et al., 2004). Yet policymakers often reject the up-front costs of effective drug treatment programs, which result in a much heavier burden on society in the long run due to incarceration, recidivism, reduced productivity, and decreased safety. Similarly, the cost to treat a person with a serious mental illness is much less than the cost of incarceration after a crime has been committed. The deinstitutionalization of mental health treatment in the 1960s and 1970s, by the shuttering of state mental hospitals, resulted in a large increase of individuals with severe mental health conditions in the U.S. prison system (Collier, 2014). While state mental hospitals may not have been a humane solution, an alternative, community-based treatment for people with serious mental illness was not developed. Thus, prisons became the new asylum for the those with severe mental health conditions. Treatment of individuals with severe mental health needs in the community focusing on medication compliance, counseling, housing support, and job opportunities is a much less expensive alternative to unsafe communities than incarcerating mentally ill individuals. The National Alliance on Mental Illness (Giliberti, 2015) estimates an annual cost of $31,000 to incarcerate an individual with mental illness, while community-based mental health care costs about $10,000 annually.

More recent data from the U.S. Department of Justice (2023) documents the Bureau of Prisons' annual cost of incarceration of an inmate in 2022 as $42,672 or $116.91 per day. A White House Issue Brief in the same year by the Council of Economic Advisors (2022) focused attention on the post-pandemic increase in mental health issues beyond the already high incidence of mental health disorders prior to the COVID-19 pandemic. Further, the brief highlighted the long-term economic burden to society of untreated mental health disorders, including costs related to unemployment, violence and crime, and incarceration. School-based mental health programs and community-based mental health services were provided as interventions with evidence of effectiveness in treating mental health and ultimately mitigating the economic costs of untreated mental health to society. Another example of a program with clear evidence of ineffectiveness, yet continued use for years despite this evidence, is the Scared Straight program. See "In the Real World" for information on the Scared Straight program and related evaluations.

Unfortunately, not all policies and programs are as well researched as investments in community-based mental health treatment and the Scared Straight program, reinforcing the need to have data regarding policy and program effectiveness. In addition, even for programs that have been well researched, such as Scared Straight, policymakers and practitioners may still decide to use them. Both of these issues, a lack of informative research as well as an underuse of available research in decision making, are of relevance to evaluators and the discipline of evaluation. Thus, I hope by the end of the text, you have the knowledge and tools to

- design rigorous and informative evaluations,
- collect evaluative information on programs and policies,
- interpret evaluation data to inform policies and programs, and
- effectively present and disseminate data to increase opportunities for use.

IN THE REAL WORLD . . .

Scared Straight was introduced in the 1970s as a program to prevent juvenile delinquency. Participants were youth at risk of becoming delinquent; the program introduced them to prisons and hardened criminals to deter them from continued criminal activity.

Multiple randomized trials in the United States showed the program did not work, and in fact was harmful to many youth (Aos et al., 2001; Lilienfeld, 2005; Petrosino et al., 2012). Youth who went through the program had a higher rate of reoffending than similar youth who did not participate in the program.

Why did policymakers continue to use the program? Because it cost less than $100 per child. It seemed like a low-risk program; if it didn't work, little money would be lost. Think again. Evaluations showed that in the long run, taxpayers and crime victims paid much more than the program costs because of the additional criminal activity of those who participated. In fact, a comprehensive cost-benefit study, which continues to be updated as the effects of Scared Straight are still being realized, estimated that for each Scared Straight participant, taxpayers and victims paid over $20,000 in net costs due to increased contact with the criminal justice system and lost labor market earnings (Washington State Institute for Public Policy, 2007, 2023).

Sources: Petrosino, A., Turpin-Petrosino, C., Hollis-Peel, M. E., & Lavenberg, J. G. (2012). Scared Straight and other juvenile awareness programs for preventing juvenile delinquency: A systematic review. *Campbell Systematic Reviews, 2013*(5). http://www.campbellcollaboration.org/media/k2/attachments/Petrosino_Scared_Straight_Update.pdf; Washington State Institute for Public Policy. (2023). *Scared Straight benefit-cost results*. http://www.wsipp.wa.gov/BenefitCost/ProgramPdf/114/Scared-Straight

1.2.1 Evaluation Is an Ethical Obligation

When presented with information regarding the costs of ineffective programs, it is not a leap to conclude that it is an ethical obligation of those who fund and implement programs and policies to also have those programs and policies evaluated. Yet program planning in general is often one of those areas where many opt to forgo evaluation, due to the up-front costs, in favor of spending those funds on program services. While serving more people may seem noble, it is not at all noble if the program is ineffective at best, and harmful at worst. While perhaps an unpopular view, my view nonetheless is that claiming ignorance to a program or policy being ineffective or even harmful, due to a lack of available data to make a judgment or due to an unwillingness to listen to available data, is an unacceptable and unethical assertion. What is your viewpoint? There is no right or wrong answer, but certainly it is an interesting question to consider.

Some readers of this text may think it is understandable that program leaders want to maximize program funds used to deliver services. I tend to agree with you, and in an ideal world, all programs would be perfectly efficient and effective allowing us to direct all funds to program services. However, that is not the case. Thus, it is a dilemma: spend money on program services (and serve more people) or spend money on evaluation (and serve fewer people). If you have donated money to a charity to provide a new after-school mentoring program for middle school students, you may want the charity to put all donated funds into the mentoring. However, what if the mentoring is ineffective and a waste of your donated dollars? Would you be willing to let the charity allocate a portion of the donation to evaluate the effectiveness of the mentoring program? Using policy and program resources to collect the necessary data to evaluate effectiveness is the only way, as Nielson might say, to live in the *light*. That is, using funds now to determine if outcomes warrant continued funding of the program in the future is a long-game mindset.

1.2.2 Evaluation Fosters Quality

It is the very nature of evaluation to increase knowledge, and this knowledge can be used to improve programs. Thus, evaluation fosters quality. It provides the necessary information to improve a program continuously, allocate resources in ways that can maximize effectiveness, and refine program strategies for greater impact. A student's improvement can be facilitated with constructive teacher feedback. An employee's performance is supported when provided with ways to improve. An organization's productivity is enhanced when there is a culture of process improvement. And the quality of a program is fostered when program components are examined and sound evaluative information is made available and utilized. Thus, the premise holds for people, organizations, and programs: When good information is provided, better decisions can be made.

Have you ever heard anyone say, "The more you know, the less you know"? This is directly tied to one of my favorite statements: Ignorance is bliss. While it might be blissful to the ignorant, to those who have to deal with the consequences of ignorance, it can be aggravating, troubling, and costly. Yet the more we learn, the more we understand all that we do not know. This journey of learning empowers us to make better, more informed decisions. And it also inspires us to search for greater understanding. Thus, evaluation produces knowledge that informs decisions, which in turn creates the need for more knowledge. This cycle of knowledge generation and use is a continuous improvement process that fosters informed decision making and, in turn, promotes quality in programs and policies.

1.2.3 Evaluation Is a Viable Career

If evaluation as an ethical obligation as well as an endeavor that fosters quality in programs and policies has not convinced you to learn all you can about evaluation, perhaps knowing that evaluation is a growing field with many job opportunities will spark your interest. Over a decade ago, Russ-Eft and Preskill (2009) included in their list of reasons to pursue a career in evaluation the increasing respect for evaluation experience as a skill that is highly marketable. Indeed, there is a need for trained evaluators within nonprofit organizations, corporations, and research centers. There are many opportunities for evaluators internationally. In this era of data-driven decision making, thankfully organizations are recognizing the value that data can provide to their operations. It is also a time of accountability, with many programs being required to show evidence of impact to receive continued funding.

1.3 VALUES AND STANDARDS IN EVALUATION

For evaluators, the creation of knowledge is based on data. But how do we decide what data to collect? How do we decide what questions to ask? And once knowledge is generated, how are the data used? In evaluation, the people who use evaluation findings are called stakeholders. In fact, a **stakeholder** is anyone who has an interest in or is involved with the operation or success of a program. Key stakeholder groups often include program staff, program participants, community members, and policymakers. In what ways do different stakeholder groups use evaluation findings? How do stakeholders weigh evaluation data in making decisions? Evaluation is the activity of examining programs and collecting information, as well as the process of determining how that information will be used. Both aspects, how we collect data and how we use data, are influenced by factors related to evaluator skills and preferences, as well as stakeholder values and the context within which the program operates.

Thus, the *valuing* that is part of evaluation is influenced by context, including our own values, the values of stakeholders, as well as politics, resources, and even history. As evaluators, it is important that we are clear about the values and standards on which our evaluative judgments are based.

An important tenet of Weiss's definition of evaluation involves the comparing of evaluation evidence to a standard, to make a judgment about a program or policy. Thus, the standard holds power. For instance, in your classes, you must achieve a certain grade to pass a course. That grade requirement is the standard. Likewise, states set cut scores for state achievement testing that are used to determine course placement and even graduation. The cut score is a standard that has the power to affect a student's future. How was that cut score set? The standard was likely set by a group of administrators based on something. That *something* likely includes data, professional judgment, research, and experience, which are all influenced by values.

Valuing is the process of estimating the importance of something. A **value** is a principle or quality that we use to estimate that importance. Value is also the estimate of importance. That is, we use values to assign value. A teacher might value effort over performance, and thus assign grades based largely on effort. A manager might value quantity of work over quality of work and use standards based on these values for employee performance appraisals.

Earlier we mentioned that evaluation, or the process of valuing, has two components: merit and worth. The merit of an evaluation might be determined by the methods used and the rigor of the design. But what influences methods and design? An evaluator's own values often guide the choice of design and methods. What does the evaluator value? Perhaps it is hearing stories from participants detailing personal experiences with the program. Or maybe the evaluator places an importance on quantitative indicators of program impact. It might be involving stakeholders in all aspects of the evaluation or possibly it is being the expert and using that expertise to design and implement the evaluation. Or it might be ensuring all possible participants receive the program being evaluated. On the other hand, maybe it is using the most rigorous evaluation design, even if that means some potential participants do not receive the program or are delayed in receiving it. There are no right answers to these questions; all are debated among evaluators.

The worth of an evaluation is dependent on context and who is making the judgment of worth. Worth to an evaluator may raise the same questions described earlier relating to merit. Worth to stakeholders is influenced by their own values. To design an evaluation that is useful to stakeholders, it is important for an evaluator to understand stakeholder values. These values will likely vary across stakeholder groups, and thus the design of the evaluation will have multiple components to address issues that allow for judgments of worth to be made.

If at this point you are ready to throw your hands up in frustration at the subjectivity inherent in valuing, instead marvel at the complexity of human thought. Okay—enough marveling. There are tools to guide evaluators in understanding not just our own values, but more important, those of our stakeholders. There are also guidelines and principles evaluators can use to conduct evaluations with objectivity.

1.3.1 Guiding Principles for Evaluators

The American Evaluation Association (AEA) provides guiding principles for evaluators. The **AEA Guiding Principles for Evaluators** (AEA, 2018a) is a set of five principles that embody the values of the AEA, an international professional association of evaluators. And yes, the guiding principles are based on values. However, these are values accepted by evaluators across disciplines and have been ratified by a membership of over 7,000 evaluators. So they have merit in their interdisciplinary nature and worth in their widespread acceptance. The guiding principles are intended to promote the ethical behavior of evaluators, and address ideals that reach across disciplinary boundaries, such as an evaluator's obligation to be professionally and culturally competent. They include guidance in the following domains:

- systematic inquiry
- competence
- integrity
- respect for people
- common good and equity

Each of these five guiding principles for evaluators is described more fully in Chapter 3. The full text of the *American Evaluation Association Guiding Principles for Evaluators* appears, with permission from the AEA, at the end of this chapter (see Figure 1.4).

1.3.2 Program Evaluation Standards

Another important resource for evaluators is a set of standards issued by the Joint Committee on Standards for Educational Evaluation. The Joint Committee is a group of representatives from multiple professional organizations, including the American Evaluation Association, that have an interest in improving evaluation quality. The **Joint Committee's Program Evaluation Standards** (Yarbrough et al., 2011) is a set of 30 standards to guide evaluators in designing and implementing quality evaluations. The standards address five areas:

- utility
- feasibility
- propriety
- accuracy
- evaluation accountability

These standards provide practical guidance on how to conduct effective and equitable evaluations that produce accurate findings and promote usability. See Figure 1.2 for a list and description of the 30 program evaluation standards.

1.3.3 Evaluator Competencies

The American Evaluation Association also provides evaluators with a set of expected competencies. The **AEA Evaluator Competencies** (AEA, 2018b) are a set of 49 competencies across five domains. They are intended to explain what it means to be an evaluator as well as create a common language that evaluators can use in referring to their practice. The five domains are:

1. professional practice
2. methodology
3. context
4. planning and management
5. interpersonal

Each of these competency domains are described more fully in Chapter 3. The full text of the *American Evaluation Evaluator Competencies* appears, with permission from the AEA, in Figure 3.3.

1.3.4 Evaluation Checklist Project

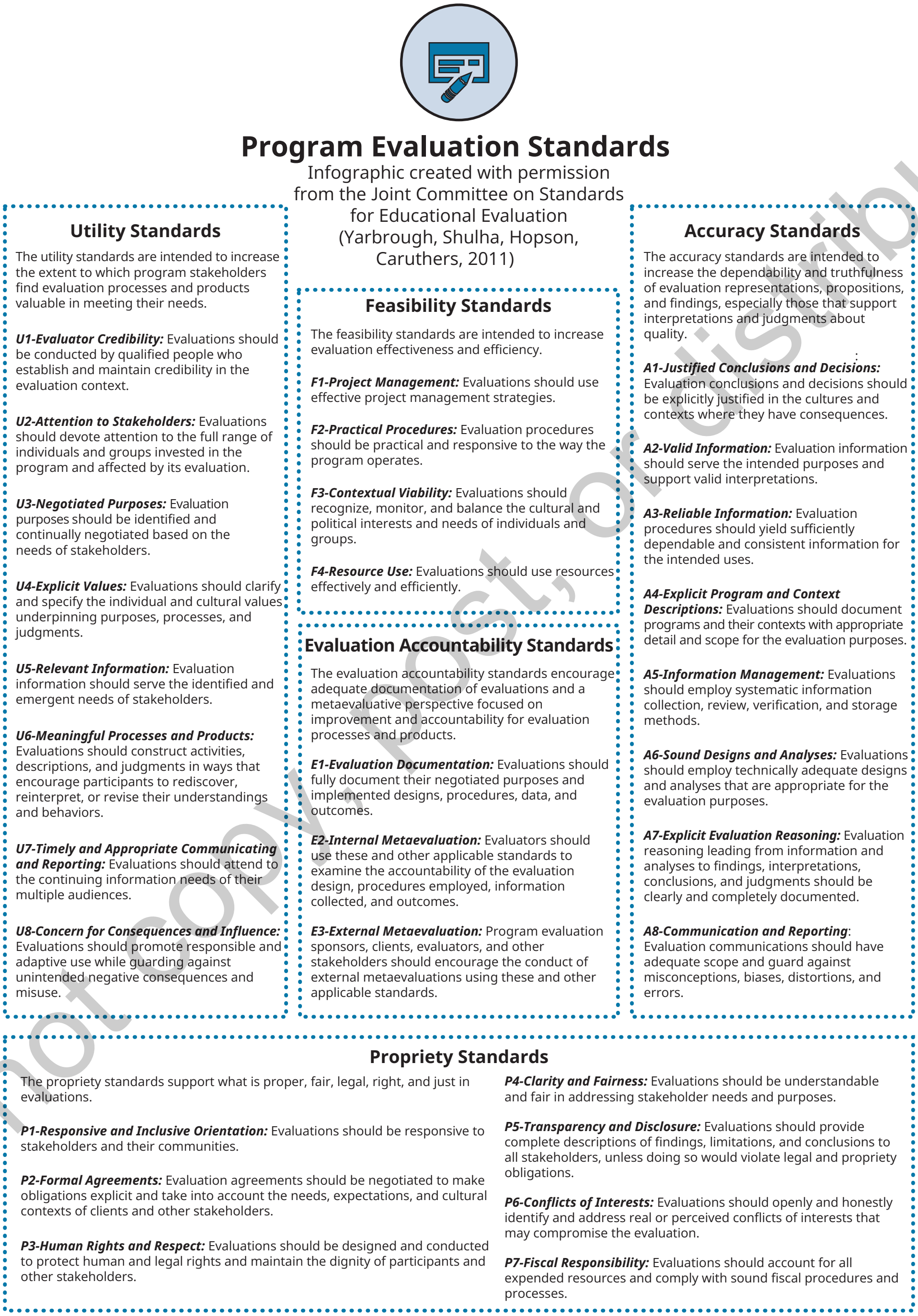
In addition to the AEA Guiding Principles, the Joint Committee's Program Evaluation Standards, and the AEA Evaluator Competencies, experienced evaluators have provided resources to guide evaluators in the appropriate consideration of values and standards in their evaluation work. The Joint Committee (2018) provides a checklist for the program evaluation standards. Stufflebeam (2001) created a checklist of values and criteria for evaluators to consider when designing and conducting evaluations. This checklist includes societal values, such as equity, effectiveness, and excellence. Also included are institutional values, such as the organization's mission, goals, and priorities. Both of these checklists are part of the Western Michigan University's Evaluation Checklist Project. This Evaluation Checklist Project was launched with funding from the National Science Foundation in 1999 and expanded in 2017 through funding from the Faster Forward Fund. These checklists are valuable resources to guide evaluators in their work. There are currently over 25 checklists, including Stufflebeam's (2004b) Evaluation Plan and Operations Checklist, Scriven's (2015) Key Evaluation Checklist, Wingate and Schroeter's (2007) Evaluation Questions Checklist, Stufflebeam's (2004a) Evaluation Design Checklist, and Feinstein's (2019) Checklist for Evaluation Recommendations. In addition to these checklists, House and Howe (1999) provide a detailed look at values in evaluation in their book *Values in Evaluation and Social Research*. Should the topic of values in evaluation spark your interest, the House and Howe text is an excellent resource through which to continue your exploration.

In summary, our own values affect all aspects of evaluation, from the research design and methods we choose to how we interact with stakeholders and the way we interpret our findings. Stakeholder values and the political context in which a program operates also affect how an evaluation is conducted and how data are used. It is important to remember that knowledge is power. Understanding stakeholder values and the political context can aid you in designing an evaluation that meets stakeholder needs and is more likely to be used to influence decision making. Understanding our own values can help us examine how they might impact our evaluations, as well as increase awareness of ways to improve our practice. Understanding and adhering to professional guidelines and standards can only serve to strengthen the work that we do as evaluators. These same professional guidelines and standards can aid stakeholders and evaluators in assessing the merit and worth of evaluation findings.

QUICK CHECK

1. What is evaluation? How do we use evaluation in our everyday lives?
2. How are research and evaluation related? What are some ways in which research and evaluation differ?
3. Why might it be considered unethical to not evaluate a program or policy?
4. How do your own values affect your views toward and choice of evaluation designs and methods?

FIGURE 1.2 ■ Joint Committee's Program Evaluation Standards



Program Evaluation Standards

Infographic created with permission
from the Joint Committee on Standards
for Educational Evaluation
(Yarbrough, Shulha, Hopson,
Caruthers, 2011)

Utility Standards

The utility standards are intended to increase the extent to which program stakeholders find evaluation processes and products valuable in meeting their needs.

U1-Evaluator Credibility: Evaluations should be conducted by qualified people who establish and maintain credibility in the evaluation context.

U2-Attention to Stakeholders: Evaluations should devote attention to the full range of individuals and groups invested in the program and affected by its evaluation.

U3-Negotiated Purposes: Evaluation purposes should be identified and continually negotiated based on the needs of stakeholders.

U4-Explicit Values: Evaluations should clarify and specify the individual and cultural values underpinning purposes, processes, and judgments.

U5-Relevant Information: Evaluation information should serve the identified and emergent needs of stakeholders.

U6-Meaningful Processes and Products: Evaluations should construct activities, descriptions, and judgments in ways that encourage participants to rediscover, reinterpret, or revise their understandings and behaviors.

U7-Timely and Appropriate Communicating and Reporting: Evaluations should attend to the continuing information needs of their multiple audiences.

U8-Concern for Consequences and Influence: Evaluations should promote responsible and adaptive use while guarding against unintended negative consequences and misuse.

Feasibility Standards

The feasibility standards are intended to increase evaluation effectiveness and efficiency.

F1-Project Management: Evaluations should use effective project management strategies.

F2-Practical Procedures: Evaluation procedures should be practical and responsive to the way the program operates.

F3-Contextual Viability: Evaluations should recognize, monitor, and balance the cultural and political interests and needs of individuals and groups.

F4-Resource Use: Evaluations should use resources effectively and efficiently.

Evaluation Accountability Standards

The evaluation accountability standards encourage adequate documentation of evaluations and a metaevaluative perspective focused on improvement and accountability for evaluation processes and products.

E1-Evaluation Documentation: Evaluations should fully document their negotiated purposes and implemented designs, procedures, data, and outcomes.

E2-Internal Metaevaluation: Evaluators should use these and other applicable standards to examine the accountability of the evaluation design, procedures employed, information collected, and outcomes.

E3-External Metaevaluation: Program evaluation sponsors, clients, evaluators, and other stakeholders should encourage the conduct of external metaevaluations using these and other applicable standards.

Accuracy Standards

The accuracy standards are intended to increase the dependability and truthfulness of evaluation representations, propositions, and findings, especially those that support interpretations and judgments about quality.

A1-Justified Conclusions and Decisions: Evaluation conclusions and decisions should be explicitly justified in the cultures and contexts where they have consequences.

A2-Valid Information: Evaluation information should serve the intended purposes and support valid interpretations.

A3-Reliable Information: Evaluation procedures should yield sufficiently dependable and consistent information for the intended uses.

A4-Explicit Program and Context Descriptions: Evaluations should document programs and their contexts with appropriate detail and scope for the evaluation purposes.

A5-Information Management: Evaluations should employ systematic information collection, review, verification, and storage methods.

A6-Sound Designs and Analyses: Evaluations should employ technically adequate designs and analyses that are appropriate for the evaluation purposes.

A7-Explicit Evaluation Reasoning: Evaluation reasoning leading from information and analyses to findings, interpretations, conclusions, and judgments should be clearly and completely documented.

A8-Communication and Reporting: Evaluation communications should have adequate scope and guard against misconceptions, biases, distortions, and errors.

Propriety Standards

The propriety standards support what is proper, fair, legal, right, and just in evaluations.

P1-Responsive and Inclusive Orientation: Evaluations should be responsive to stakeholders and their communities.

P2-Formal Agreements: Evaluation agreements should be negotiated to make obligations explicit and take into account the needs, expectations, and cultural contexts of clients and other stakeholders.

P3-Human Rights and Respect: Evaluations should be designed and conducted to protect human and legal rights and maintain the dignity of participants and other stakeholders.

P4-Clarity and Fairness: Evaluations should be understandable and fair in addressing stakeholder needs and purposes.

P5-Transparency and Disclosure: Evaluations should provide complete descriptions of findings, limitations, and conclusions to all stakeholders, unless doing so would violate legal and propriety obligations.

P6-Conflicts of Interests: Evaluations should openly and honestly identify and address real or perceived conflicts of interests that may compromise the evaluation.

P7-Fiscal Responsibility: Evaluations should account for all expended resources and comply with sound fiscal procedures and processes.

1.4 TYPES OF EVALUATION

The terms evaluation, program, and research were explained in the previous section. There are several additional important evaluation terms that are explained in this section. The terminology introduced in this section is often used in evaluation solicitations and requests for proposals. A **request for proposal (RFP)** is an announcement that an agency has funds available for specified work and an invitation for organizations to prepare a description of how they would complete that work. Some RFPs ask specifically for evaluation services and some may ask for program development or implementation, with a stipulation that an evaluation plan must be included in the proposal. RFPs will often use language indicating that formative and summative evaluation is required, or an external evaluator is preferred. Some organizations use the term request for applications (RFA) to announce opportunities for funding. The federal government uses the terms **notice of funding opportunity (NOFO)** or funding opportunity announcement (FOA) to publicly announce the availability of grant funding opportunities.

RFPs and NOFOs are two of many methods through which you might hear about the need for an evaluation. While some of the evaluations I have conducted originated with an RFP or NOFO, most of my evaluation work comes when an individual or organization directly contacts our evaluation center. Sometimes we are asked to evaluate a program that is being planned, and sometimes we are invited to evaluate a program already in operation. Other times we are asked to write an evaluation plan for a project being proposed and submitted for funding, with the understanding that should the project be funded, we will conduct the evaluation. Evaluators might also be hired to be part of an organization; some larger organizations have evaluators on staff to conduct routine evaluations of their programs and policies. Regardless of whether an evaluation comes about due to an RFP or NOFO, direct contact, in-house planning, or some other means, the terms presented in this section are commonly used when requesting evaluation assistance.

The framework presented in this section to introduce evaluation terminology is adapted from Trochim's (2001, 2016) *The Research Methods Knowledge Base*. He categorizes common types of evaluation within the formative and summative domains. **Formative evaluation** is evaluation aimed at providing information to program staff so they can improve the program while it is in operation; formative evaluation methods include process evaluation, implementation assessment, needs assessment, and evaluability assessment. **Summative evaluation** is evaluation aimed at providing information to program staff regarding effectiveness so they can make decisions about whether to continue or discontinue a program; summative evaluation methods include outcome evaluation, impact evaluation, cost-effectiveness/cost-benefit analysis, and meta-analysis. Many evaluations have both formative and summative components, with the formative component geared toward increasing the impact measured by the summative evaluation.

1.4.1 Formative Evaluation

An important purpose of evaluation is to collect information that enables program staff to improve a program while it is in operation and prepares program leadership to make decisions about a program's future. This is called formative evaluation. At the root of the word "formative" is *form*. To form is to shape. Thus, formative decisions are those that are intended to form, shape, and improve a program while being implemented. Formative evaluation is performed to provide ongoing data to program staff for continuous improvement. Formative evaluation examines the implementation process, as well as outcomes measured throughout program implementation, to make decisions about midcourse adjustments, technical assistance, or professional development that may be needed, as well as to document the program's implementation so that others can learn from the program's operation.

Evaluators use **process evaluation** to make midcourse adjustments to shape a program. When evaluators conduct a process evaluation, they examine the output of the process of implementing a program's operations. A process evaluation might focus on the number of people trained, types of services delivered, methods of training used, and other measures related to delivering program services.

Another type of formative evaluation is **implementation assessment**, that is, determining the degree to which a program is implemented as planned. Implementation assessment examines the fidelity with which a program's strategies or activities have been implemented. To assess fidelity of implementation, one must have a model of how a program would "look" if it was implemented as envisioned. Likewise, having a sense of how a program implementation is not optimal can help establish the degree of fidelity. For instance, think about how your teachers might use a rubric to grade a written assignment. Let's suppose your rubric scores range from 0 to 10 on multiple components of the paper, for example, identification of thesis, organization, and grammar. The rubric would tell you what it means to get a 10 on organization versus a 5 or 0. That way, you would know what the teacher sees as an "ideal" paper versus an average or below average paper. Similarly, implementation assessment can help evaluators understand how various activities within a program are implemented and the degree to which implementation matches the intentions of the program developers.

Prior to implementing a new program or restructuring an existing program, **needs assessment** can be used to shape the program by examining the needs of proposed participants, needs of stakeholders, and how to meet the needs of both. Needs assessment is a systematic examination of what program services are needed, who needs these services, and in what ways they need the services.

Finally, **evaluability assessment** helps determine whether it is feasible to conduct an evaluation of a particular program (Trevisan, 2007; Wholey, 1979, 2002). It addresses whether a program or policy has clearly defined outcomes of interest; if it is feasible to attribute outcomes to the program or policy; whether data are available, reliable, and valid; whether stakeholders are identifiable and accessible; if the necessary resources are available to conduct the evaluation; and the likelihood that findings will be used appropriately. Evaluability assessment also examines how stakeholders might be involved within the evaluation to shape the program and its attendant evaluation in a way that best meets the determined needs. An excellent resource on how to conduct evaluability assessments is *Evaluability Assessment: Improving Evaluation Quality and Use* by Trevisan and Walser (2015).

1.4.2 Summative Evaluation

A primary purpose of evaluation is to make summative decisions. Summative decisions are made by looking at all the information available regarding the program. At the root of the word "summative" is *sum*. A "sum" is a total or a result. Thus, summative evaluation is performed to make final, outcome-related decisions about program funding. Summative decisions include whether to continue, expand, or discontinue a program based on evaluation findings. Formative evaluation findings can be used to prepare program leadership to make such decisions. Formative evaluation findings can also be used to improve programs so that the program has increased opportunity to positively affect summative findings.

Summative evaluation speaks to decisions about a program's future. As such, **outcome evaluation** is summative evaluation focused on how well a program met its specified long-term goals. If a program proposes to improve learning, outcome evaluation would focus on changes in knowledge. If a program proposes to change practices related to healthy eating or medication compliance, outcome evaluation would focus on behavior change.

Impact evaluation also measures the outcomes of programs. However, **impact evaluation** is broader than outcome evaluation, as it measures all impacts of a program, both those intended as specified by a program's goals and those unintended. For example, an impact evaluation of No Child Left Behind (NCLB) showed that principals and teachers made better use of test data after NCLB was passed and that scores on state tests had increased (Center on Education Policy, 2006). However, the study also showed that the curriculum had narrowed to focus on tested material, student creativity had declined, and flexibility within the law might account for more students being classified as proficient (Center on Education Policy, 2006). Other studies have shown that NCLB decreased the average quality of principals at disadvantaged schools due to principals seeking employment at schools less likely to experience NCLB sanctions (Li, 2010), reduced educational programming for gifted students (Beisser, 2008), and raised new challenges specific to using accommodations in high-stakes testing (Cawthon, 2007).

Many program funders request information on the efficiency of a program. That is, they want to know the value of a program, either in terms of dollars saved or outcomes measured, or in its benefits to participants or society. Hence, **cost-benefit/cost-effectiveness analysis (CBA/CEA)** is summative evaluation that focuses on estimating the efficiency of a program regarding dollar costs saved (cost-benefit) or outcomes observed (cost-effectiveness). The amount saved by a program might differ depending on the time frame used for the analysis. For instance, recall the example about mental health treatment at the start of the chapter. Estimating the amount saved 1 year out would not give a full picture of the benefits of the program; the cost savings for some programs are not realized until years or decades after the program ends. The allure of funded preschool programs is not simply preparing a child for kindergarten, but rather, proponents argue, the long-term benefits of preschool programs include increased high school graduation rates, which in turn lead to increased employability and improved quality of life. Estimating the cost-benefit of a program intended to have long-term cost savings is difficult, but can be done (see "In the Real World" on the Scared Straight program). An alternative to measuring program success regarding cost savings is calculating the benefits of a program concerning nonmonetary outcomes. While cost-benefit is a ratio of the costs of the program to the costs saved by the program, cost-effectiveness is calculated by using a ratio of the total costs associated with program delivery to the impact of the program on a chosen outcome. For instance, a behavioral intervention program might measure cost-effectiveness as the change in behavioral outcomes for every $1 spent on the program. A program targeting healthy eating might estimate the change in fast food consumption per dollar spent on the program or weight loss associated with each dollar invested in the program. For more information, see Cellini and Kee's (2015) chapter on cost-effectiveness and cost-benefit analysis in the *Handbook of Practical Program Evaluation* (Newcomer et al., 2015). **Cost-utility analysis** (CUA) is a type of cost-effectiveness analysis; it is an economic analysis often used with health interventions to examine the cost of a program in relation to quality of life improvements. These improvements may be defined in terms of extension of life in years or number of lives saved.

Meta-analysis is a form of summative evaluation that integrates the findings of multiple studies to estimate the overall effect of a program or type of program. Meta-analysis is a statistical approach that merges results across a body of research. Such analyses are also referred to as systematic reviews because the methodology is highly structured and involves defining inclusion and exclusion criteria for prospective studies, combining measures across studies, and calculating new estimates of effectiveness based on the pooled data. See the What Works Clearinghouse (https://ies.ed.gov/ncee/wwc/) for more information on systematic reviews of evidence, as well as their Tier 1, Tier 2, and Tier 3 rating system.

Finally, a **meta-evaluation** is an evaluation of an evaluation (Scriven, 1969, 2009; Stufflebeam, 1978). Formative meta-evaluations provide feedback to improve an evaluation. These evaluations might be internal (internal formative meta-valuation; IFME) or conducted by an external evaluator (external formative meta-evaluations; EFME). Summative meta-evaluations assess the quality and merits of an evaluation. The Joint Committee's Program Evaluation Standards, discussed earlier, can be used to conduct meta-evaluations.

This text focuses primarily on the process evaluation and implementation assessment components of formative evaluation and the outcome and impact evaluation components of summative evaluation. However, resources are provided in subsequent chapters for additional information on needs assessment, evaluability assessment, cost-benefit/cost- effectiveness analysis, meta-analysis, and meta-evaluation.

1.5 INTERNAL AND EXTERNAL EVALUATION

Because internal and external evaluation are terms commonly used by those within and outside of the evaluation field, I include them under types of evaluation. However, it should be noted that they are not types of evaluation like formative and summative evaluation, but rather a way of describing the relationship of the evaluator to the program itself.

An evaluation can be conducted by someone inside the organization within which a program operates or by someone outside of the organization. However, the optimal arrangement is often a partnership between the two, that is, forming an evaluation team that includes both internal and external evaluators.

An **internal evaluator** may be someone at the organization who is knowledgeable about the program. For evaluations that focus on program improvement and effectiveness, having an internal evaluator on the evaluation team can cultivate a deeper understanding of the context in which the program operates. Involving people inside the organization also helps build capacity within the organization to conduct evaluations. Due to their relationship to the organization, an internal evaluator may also be perceived as less threatening than an external evaluator, which might foster greater evaluation buy-in from program staff and an increased likelihood that evaluation findings will be used by the organization. However, an internal evaluator should be someone who is in a position to be objective regarding program strengths and weaknesses. For this reason, choosing an internal evaluator who is *responsible* for the program's success is not recommended and may compromise the evaluation. Likewise, any time an internal evaluator is very close to the program being evaluated, objectivity or perceived objectivity may suffer. To maintain objectivity, an internal evaluator should be outside of the program. However, while staff from within the program may not be part of the core evaluation team, they should certainly partner with the evaluation team to ensure that the evaluation informs the program during every phase of implementation.

An **external evaluator** is an evaluator who is employed from outside of the organization that operates the program or policy to be evaluated. It is good practice to have an external evaluator be part of your evaluation team. Using an external evaluator as a "critical friend" provides you with an extra set of eyes and a fresh perspective from which to review your design and results. Professional evaluators are trained in the design of evaluations to improve usability of the findings, and they are skilled in data collection techniques such as designing surveys, facilitating focus groups, conducting interviews, choosing quality assessments, and performing observations. An experienced evaluator can also help you analyze and interpret your data, as well as guide you in the use of your results.

Partnering with an external evaluator can improve the **credibility** of the findings, as some may question whether an evaluator from within an organization can have the **objectivity** to recognize areas for improvement and to report results that might be unfavorable to the program. This is not to imply that credibility or objectivity problems are usual or even common with internal evaluations. External as well as internal evaluations can suffer from a lack of credibility or objectivity. But issues of credibility and objectivity in internal evaluations arise due to a perceived threat to the findings. For that reason, it is important for evaluators to disclose, in a straightforward manner, any conflicts of interest or connections to the program under evaluation when reporting evaluation findings.

Note that an external evaluator may be a researcher or professor from your local university, a professional evaluator from a private evaluation firm, or an independent evaluation consultant. For programs where an external evaluator might be preferred, funding an outside evaluator may not be feasible. In such cases, partnering with an evaluator within your organization, yet outside of your program, might work well. For instance, when evaluating education programs, staff from a curriculum and instruction office implementing a program might partner with staff from another office within the school district, such as an assessment or evaluation office, to conduct the evaluation.

If resources are not available for an external evaluator and there is no office or department in your organization that is not affected by your program, you may want to consider other potentially affordable evaluation options. You could put out a call to individuals with evaluation experience within your community who might be willing to donate time to your program; contact a local university or community college regarding faculty or staff with evaluation experience who might work with you at a reduced rate; ask your local university if there is a doctoral student in evaluation who is looking for a research opportunity or dissertation project; or explore grant opportunities that fund evaluation activities.

The choice of who conducts your evaluation should depend on the anticipated use of the results and the intended audience, as well as your available resources. If evaluation results are to be used with current or potential funding agencies to secure support and assistance, contracting with an external evaluator would be your most prudent choice. If the evaluation is primarily intended for use by your organization to improve programs and understand impact, an evaluation team composed of an internal and an external evaluator may be preferable. Connecting with someone outside your organization to assist with the evaluation and results interpretation will likely enhance the usability of your evaluation and the credibility of your evaluation findings. Evaluation as a partnership between an internal evaluator and an external evaluator is the ideal arrangement to ensure the utility of the evaluation and its results.

I would like to add that there is no established, best practices framework for how internal and external evaluators collaborate to evaluate a program. This is an area in which we are still learning what relationship is most effective and efficient. Some internal-external evaluator models keep the two roles clearly delineated and separate, notwithstanding a once or twice a year communication around evaluation reporting. Other models blend the internal and external evaluator roles to maximize the efficiency of resources and to leverage the particular skillset of each evaluator. I have evaluated programs using both of these models, as well as other models somewhere in between these two extremes. The latter framework, internal and external evaluators working together as a team to benefit the program, requires a degree of trust between evaluators and often a working relationship built on previous collaborations. On the other hand, maintaining separation and independence between the internal and external evaluator roles may create redundancy in data collection and complicate the understanding of findings should they not be consistent across evaluators. To guard against redundancy in such evaluations, a strategy I have found effective situates the internal evaluator as conducting the day-to-day evaluation and the external evaluator conducting a meta-evaluation of the internal evaluation annually or every other year. Regardless of the model used, communication between an internal and external evaluator, an understanding by program leadership and evaluators of how evaluation roles are defined, and support from the funding agency regarding degree of collaboration between the internal and external evaluator are critical to the successful evaluation of a program and the usefulness of evaluation findings. Perhaps as the discipline of evaluation continues to grow and as evaluators share with one another what works well and what does not, best practices regarding evaluator collaboration for particular types of programs, different program settings, or intended use of evaluation findings will emerge.

Overall, it is important to remember that both internal and external evaluations have their benefits and drawbacks. In determining the structure of who conducts an evaluation, weigh the extent to which *perceived* objectivity is a threat to evaluation credibility, as well as the ways in which different stakeholder groups might use the findings.

QUICK CHECK

1. What is formative evaluation? How does formative evaluation differ from summative evaluation?
2. How can implementation assessment be used to make formative and summative evaluation decisions?
3. Why might someone be skeptical of an evaluation conducted by an internal evaluator? What can be done to strengthen perceived objectivity when an internal evaluator is used?
4. How might internal and external evaluators work together to benefit a program?

1.6 EMBEDDING EVALUATION INTO PROGRAMS

The resource tug-of-war between program services and program evaluation is real and has implications for the short- and long-term implementation of program. It is this dilemma that has shaped the way in which I work with my clients so that evaluation is useful in not only determining the outcomes of their programs but also in helping improve their programs on an ongoing basis. Embedded evaluation is an evaluation approach that can be built into programs and processes so that it is part of everyday practice. This method recognizes the preciousness of resources and time, the need for information, and the tension between the two. The embedded approach to evaluation is not an additional step to be superimposed on a program and the strategies it employs, but rather a way to weave evaluation into the design, development, and implementation of policies, programs, and projects.

1.6.1 Grounded in Continuous Improvement

Embedded evaluation incorporates the underlying philosophies of both total quality management (TQM) and quality improvement (QI) initiatives, in that the purpose of embedding evaluation into your programs is to create continuous improvement processes and improve the quality and utility of data collected. Thus, **embedded evaluation** is a method of continuous improvement in which processes and practices are examined and refined to improve outcomes.

If you are not familiar with TQM or QI, TQM is a philosophy and an approach used to improve processes within organizations. See the American Society for Quality (ASQ) website for more information on TQM (asq.org). TQM is based on quality improvement principles.

QI is concerned with improving performance in a systematic and continuous manner. Its processes are also referred to as continuous improvement (CI). The U.S. Department of Health and Human Services (Health Resources and Services Administration, 2011) has made available a resource on the principles and processes of QI. This report, titled *Quality Improvement*, explains QI and provides practical guidance in creating and implementing QI programs.

1.6.2 Theory Based and Utilization Focused

The embedded evaluation approach presented in this textbook is one of many approaches that can be used when conducting an evaluation (note that Chapter 4 provides a comprehensive review of evaluation approaches). Embedded evaluation combines elements from several common evaluation approaches, including theory-based evaluation, stakeholder evaluation, participatory evaluation, and utilization-focused evaluation. Theory-based evaluation, in particular, focuses on indicators related to the logic underlying a program to guide evaluation. Utilization-focused evaluation is based on the premise that an evaluation's worth rests in how useful it is to the program's stakeholders. Both theory-based and utilization-focused evaluation approaches, as well as stakeholder and participatory evaluation, are described in detail in Chapter 4.

1.6.3 Dynamic and Cyclical

Earlier in this chapter I mentioned that evaluation is a lot like the scientific method: You define a problem, investigate the problem, document results, refine the problem based on lessons learned from the results, investigate again, and so on. Although the steps of embedded evaluation presented in this text may appear as if they are linear rungs on a ladder culminating with the final step, they are not rigid steps. Rather, embedded evaluation steps build on each other and depend on decisions made in prior steps, and information learned in one step may lead to refinement of another step. Similar to the scientific method, embedded evaluation is cyclical. The steps of embedded evaluation are components of the evaluation process that impact and influence each other. What you learn or decide in one step may prompt you to return to a previous step for modification and improvement. Just as programs are ongoing, evaluation is dynamic.

The dynamic nature of evaluation and the interconnectedness of an embedded evaluation with the program itself may seem amiss to researchers who prefer to study a phenomenon over time and wait until a predefined time to analyze and report findings. And inarguably, having a program stay its course without midcourse refinements and improvements would make cross-site comparisons and replication easier. To reiterate, embedded evaluation is built on the principle of continuous program improvement. With embedded evaluation, as information is gathered and lessons are learned, the program is improved. However, embedded evaluation goes beyond simply program monitoring. It is a way to build evaluation into a program, as well as to monitor implementation and assess effectiveness.

The focus of embedded evaluation is to enable program staff to build and implement high-quality programs that are continuously improving, as well as to determine when programs are not working and need to be discontinued. The overall purpose of designing a rigorous, embedded evaluation is to aid program staff in providing effective services to their clients.

1.6.4 Program Specific

Just as the first step in solving a problem is to understand the problem, the first step in conducting an evaluation is to *understand what you want to evaluate*. For the purposes of this textbook, what you want to evaluate, the evaluand, is referred to as the "program." As noted earlier, the term "program" is used broadly throughout this textbook to represent small interventions, groups of activities, community-based services, agencywide projects, and statewide initiatives, as well as national or international policy.

You can use the evaluation process presented in this textbook to define and evaluate a small project, as well as to understand and evaluate the inner workings of large programs and initiatives. Regardless of the size or type of program, understanding the program is not only the first step in evaluation; it is the most important step. Defining why a program should work and making the theory that underlies a program explicit lay the foundation upon which you can foster program improvement and measure program effectiveness. Further, understanding the program enables you to develop evaluation questions and define metrics, in collaboration with stakeholders, that are meaningful and useful to stakeholders. Understanding how the program operates can also aid you in integrating processes for the collection and use of these indicators into everyday program operations.

1.6.5 A Framework for Evaluation

Embedded evaluation is a framework grounded in continuous improvement, based on a program's theory, focused on utilizing results, dynamic and cyclical in its operation, and built into a specific program's operations to foster data-driven decision making. Chapters 5–12 guide you through designing and conducting an evaluation using this framework. You are led step-by-step from documenting how and why a program works to using evaluation results. The embedded evaluation framework is presented graphically in Figure 1.3. The framework is based on the following five steps:

Step 1. DEFINE: What is the program? (Chapters 5–6)

Step 2. PLAN: How do I plan the evaluation? (Chapters 7–8)

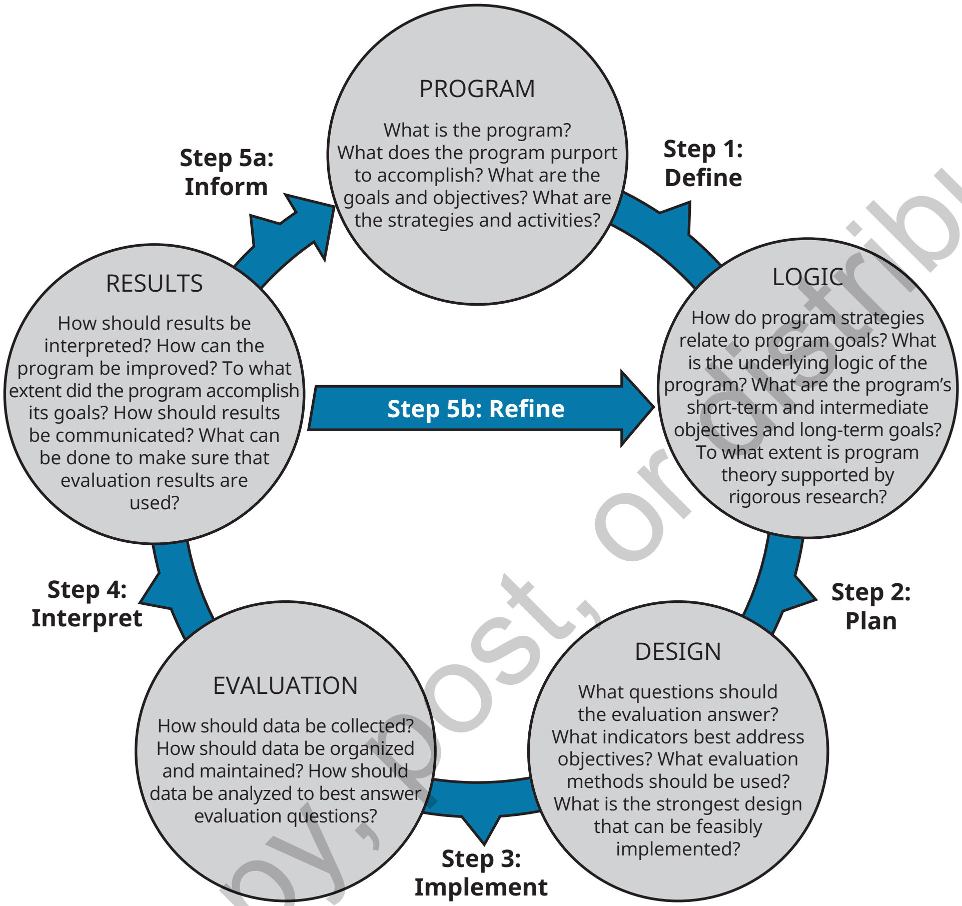
Step 3. IMPLEMENT: How do I evaluate the program? (Chapters 9–10)

Step 4. INTERPRET: How do I interpret the results? (Chapter 11)

Step 5. (a) INFORM and (b) REFINE: How do I use the results? (Chapter 12)

Prior to embarking on the embedded evaluation process, Chapters 2–4 provide a necessary foundation for future evaluators. This foundation includes a contextual understanding of the history of evaluation and its development over time; an awareness of the ethical obligations of an evaluator and the history of ethical abuses that make this awareness necessary; and a conceptual understanding of different approaches to evaluation.

FIGURE 1.3 ■ The Embedded Evaluation Model



PROGRAM

What is the program? What does the program purport to accomplish? What are the goals and objectives? What are the strategies and activities?

LOGIC

How do program strategies relate to program goals? What is the underlying logic of the program? What are the program's short-term and intermediate objectives and long-term goals? To what extent is program theory supported by rigorous research?

DESIGN

What questions should the evaluation answer? What indicators best address objectives? What evaluation methods should be used? What is the strongest design that can be feasibly implemented?

EVALUATION

How should data be collected? How should data be organized and maintained? How should data be analyzed to best answer evaluation questions?

RESULTS

How should results be interpreted? How can the program be improved? To what extent did the program accomplish its goals? How should results be communicated? What can be done to make sure that evaluation results are used?

Whether the program you are evaluating is a new program or one that has been in operation for many years, the process of embedding evaluation into your program is the same. Explicitly defining the program is a critical first step toward responsible program management, as well as program improvement. Program staff and program evaluators should have a clear and shared understanding of the program and its intended goals, as well as how and why the strategies that the program employs relate to the program's goals.

1.6.5.1 Embedding Evaluation Into Program Development

For a new program, embedding evaluation into the program development process allows data to be built into all future decision making for the program. It provides the opportunity for information to be the foundation of the program's operation from day one. Embedding evaluation during program development also provides the most flexibility with evaluation design, often allowing a more rigorous evaluation than may be feasible with an existing program. When evaluators are involved from the very beginning of a program's development, it provides an opportunity for the evaluator to work collaboratively with program staff to integrate evaluation into the program's operation and to build capacity within the program itself to use and rely on data for decision making.

1.6.5.2 Embedding Evaluation Into Existing Programs

Existing programs with good documentation and established data management systems may find embedding evaluation into the program a relatively straightforward and educational process. Existing programs with poor documentation and little data supporting their operations may find the process similar to that of embedding evaluation into a new program.

Taking the time to document the logic of an existing program can not only clarify all aspects of a program's implementation, but also provide a good opportunity for program leadership and evaluators to co-examine existing strategies and their relation to the program's goals. The process of embedding evaluation into existing programs can also aid in developing a common understanding of program goals and help foster buy-in among stakeholders. While examining the program's logic, you will likely uncover data needs that must be adopted by the program. Fostering broad stakeholder involvement during the embedded evaluation process often makes any additional data collection needs easier to implement. However, even if changing data collection methods is cumbersome, remember, it is the responsibility of an evaluator to provide program staff with the information necessary to determine how well a program working for its participants and how data-based program changes might foster improvement. Just as any organization must periodically reexamine and reaffirm its mission, all programs should routinely examine the logic underlying the program and refine that logic as necessary as lessons are learned and results are measured.

1.7 TEXTBOOK ORGANIZATION

This textbook provides you with the tools to embed evaluation into programs to foster continuous improvement, by making information and data the basis upon which the program operates. The textbook is divided into four sections:

- Section 1 includes general evaluation background, including information on key terms (Chapter 1), the history of evaluation (Chapter 2), ethical considerations in evaluation (Chapter 3), and evaluation approaches (Chapter 4).

QUICK CHECK

1. Embedded evaluation is
   a. an evaluation approach used to continuously improve the program.
   b. an evaluation approach that focuses solely on the program's staff.
   c. a linear approach to evaluation.

2. The first step in evaluation is to
   - **a.** collect data about the program.
   - **b.** decide on the methods to be used.
   - **c.** understand the program.
3. What common method is the embedded evaluation approach similar to? In what ways is it similar to this method?

Answers: 1-a; 2-c; 3-scientific method

FIGURE 1.4 ■ AEA Guiding Principles for Evaluators

Glossary

Common Good — the shared benefit for all or most members of society including equitable opportunities and outcomes that are achieved through citizenship and collective action. The common good includes cultural, social, economic, and political resources as well as natural resources involving shared materials such as air, water and a habitable earth.

Contextual Factors — geographic location and conditions; political, technological, environmental, and social climate; cultures; economic and historical conditions; language, customs, local norms, and practices; timing; and other factors that may influence an evaluation process or its findings.

Culturally Competent Evaluator — "[an evaluator who] draws upon a wide range of evaluation theories and methods to design and carry out an evaluation that is optimally matched to the context. In constructing a model or theory of how the evaluand operates, the evaluator reflects the diverse values and perspectives of key stakeholder groups."[1]

Environment — the surroundings or conditions in which a being lives or operates; the setting or conditions in which a particular activity occurs.

Equity — the condition of fair and just opportunities for all people to participate and thrive in society regardles of individual or group identity or difference. Striving to achieve equity includes mitigating historic disadvantage and existing structural inequalities.

Guiding Principles vs. Evaluation Standards — the Guiding Principles pertain to the ethical conduct of the evaluator whereas the Evaluation Standards pertain to the quality of the evaluation.

People or Groups — those who may be affected by an evaluation including, but not limited to, those defined by race, ethnicity, religion, gender, income, status, health, ability, power, underrepresentation, and/or disenfranchisement.

Professional Judgment — decisions or conclusions based on ethical principles and professional standards for evidence and argumentation in the conduct of an evaluation.

Stakeholders — individuals, groups, or organizations served by, or with a legitimate interest in, an evaluation including those who might be affected by an evaluation.

[1] The quotation is from the "Public Statement on Cultural Competence in Evaluation." Washington DC: Author. p. 3.

Purpose of the Guiding Principles: The Guiding Principles reflect the core values of the American Evaluation Association (AEA) and are intended as a guide to the professional ethical conduct of evaluators.

Focus and Interconnection of the Principles: The five Principles address systematic inquiry, competence, integrity, respect for people, and common good and equity. The Principles are interdependent and interconnected. At times, they may even conflict with one another. Therefore, evaluators should carefully examine how they justify professional actions.

Use of Principles: The Principles govern the behavior of evaluators in all stages of the evaluation from the initial discussion of focus and purpose, through design, implementation, reporting, and ultimately the use of the evaluation.

Communication of Principles: It is primarily the evaluator's responsibility to initiate discussion and clarification of ethical matters with relevant parties to the evaluation. The Principles can be used to communicate to clients and other stakeholders what they can expect in terms of the professional ethical behavior of an evaluator.

Professional Development about Principles: Evaluators are responsible for undertaking professional development to learn to engage in sound ethical reasoning. Evaluators are also encouraged to consult with colleagues on how best to identify and address ethical issues.

Structure of the Principles: Each Principle is accompanied by several sub-statements to amplify the meaning of the overarching principle and to provide guidance for its application. These sub-statements do not include all possible applications of that principle, nor are they rules that provide the basis for sanctioning violators. The Principles are distinct from Evaluation Standards and evaluator competencies.

Evolution of Principles: The Principles are part of an evolving process of self-examination by the profession in the context of a rapidly changing world. They have been periodically revised since their first adoption in 1994. Once adopted by the membership, they become the official position of AEA on these matters and supersede previous versions. It is the policy of AEA to review the Principles at least every five years, engaging members in the process. These Principles are not intended to replace principles supported by other disciplines or associations in which evaluators participate.

(Continued)

FIGURE 1.4 ■ AEA Guiding Principles for Evaluators (*Continued*)

A: Systematic Inquiry: Evaluators conduct data-based inquiries that are thorough, methodical, and contextually relevant.

A1. Adhere to the highest technical standards appropriate to the methods being used while attending to the evaluation's scale and available resources.
A2. Explore with primary stakeholders the limitations and strengths of the core evaluation questions and the approaches that might be used for answering those questions.
A3. Communicate methods and approaches accurately, and in sufficient detail, to allow others to understand, interpret, and critique the work.
A4. Make clear the limitations of the evaluation and its results.
A5. Discuss in contextually appropriate ways the values, assumptions, theories, methods, results, and analyses that significantly affect the evaluator's interpretation of the findings.
A6. Carefully consider the ethical implications of the use of emerging technologies in evaluation practice.

B: Competence: Evaluators provide skilled professional services to stakeholders.

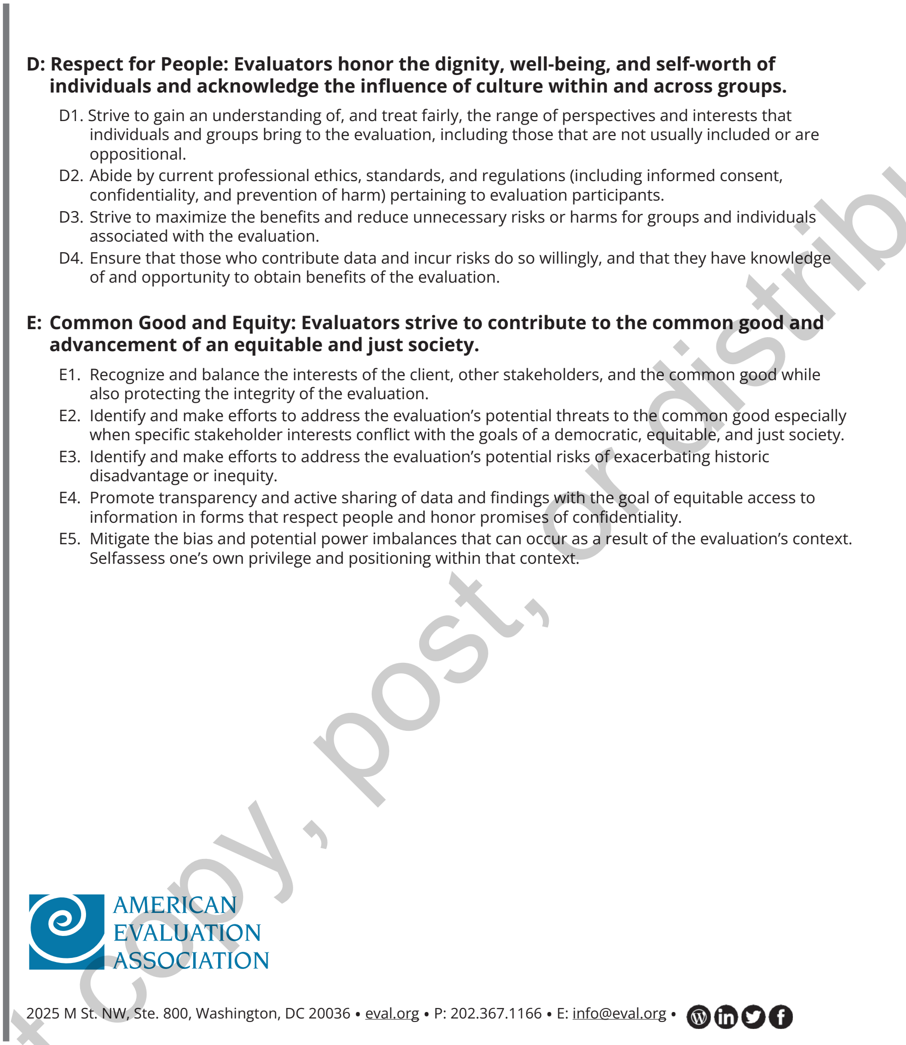
B1. Ensure that the evaluation team possesses the education, abilities, skills, and experiences required to complete the evaluation competently.
B2. When the most ethical option is to proceed with a commission or request outside the boundaries of the evaluation team's professional preparation and competence, clearly communicate any significant limitations to the evaluation that might result. Make every effort to supplement missing or weak competencies directly or through the assistance of others.
B3. Ensure that the evaluation team collectively possesses or seeks out the competencies necessary to work in the cultural context of the evaluation.
B4. Continually undertake relevant education, training or supervised practice to learn new concepts, techniques, skills, and services necessary for competent evaluation practice. Ongoing professional development might include formal coursework and workshops, self-study, self- or externally-commissioned evaluations of one's own practice, and working with other evaluators to learn and refine evaluative skills and expertise.

C: Integrity: Evaluators behave with honesty and transparency in order to ensure the integrity of the evaluation.

C1. Communicate truthfully and openly with clients and relevant stakeholders concerning all aspects of the evaluation, including its limitations.
C2. Disclose any conflicts of interest (or appearance of a conflict) prior to accepting an evaluation assignment and manage or mitigate any conflicts during the evaluation.
C3. Record and promptly communicate any changes to the originally negotiated evaluation plans, the rationale for those changes, and the potential impacts on the evaluation's scope and results.
C4. Assess and make explicit the stakeholders', clients', and evaluators' values, perspectives, and interests concerning the conduct and outcome of the evaluation.
C5. Accurately and transparently represent evaluation procedures, data, and findings.
C6. Clearly communicate, justify, and address concerns related to procedures or activities that are likely to produce misleading evaluative information or conclusions. Consult colleagues for suggestions on proper ways to proceed if concerns cannot be resolved, and decline the evaluation when necessary.
C7. Disclose all sources of financial support for an evaluation, and the source of the request for the evaluation.

(*Continued*)

FIGURE 1.4 ■ AEA Guiding Principles for Evaluators (*Continued*)



D: Respect for People: Evaluators honor the dignity, well-being, and self-worth of individuals and acknowledge the influence of culture within and across groups.

D1. Strive to gain an understanding of, and treat fairly, the range of perspectives and interests that individuals and groups bring to the evaluation, including those that are not usually included or are oppositional.

D2. Abide by current professional ethics, standards, and regulations (including informed consent, confidentiality, and prevention of harm) pertaining to evaluation participants.

D3. Strive to maximize the benefits and reduce unnecessary risks or harms for groups and individuals associated with the evaluation.

D4. Ensure that those who contribute data and incur risks do so willingly, and that they have knowledge of and opportunity to obtain benefits of the evaluation.

E: Common Good and Equity: Evaluators strive to contribute to the common good and advancement of an equitable and just society.

E1. Recognize and balance the interests of the client, other stakeholders, and the common good while also protecting the integrity of the evaluation.

E2. Identify and make efforts to address the evaluation's potential threats to the common good especially when specific stakeholder interests conflict with the goals of a democratic, equitable, and just society.

E3. Identify and make efforts to address the evaluation's potential risks of exacerbating historic disadvantage or inequity.

E4. Promote transparency and active sharing of data and findings with the goal of equitable access to information in forms that respect people and honor promises of confidentiality.

E5. Mitigate the bias and potential power imbalances that can occur as a result of the evaluation's context. Selfassess one's own privilege and positioning within that context.

AMERICAN
EVALUATION
ASSOCIATION

2025 M St. NW, Ste. 800, Washington, DC 20036 • eval.org • P: 202.367.1166 • E: info@eval.org •

- Section 2 includes a step-by-step approach to designing an embedded evaluation. It is not intended to be simply a "how to" lesson but rather a comprehensive approach to support you in planning and understanding programs, with a rigorous evaluation included as an integral part of the program's design. The section includes understanding the program (Chapter 5), modeling the program (Chapter 6), planning the evaluation (Chapter 7), and designing the evaluation (Chapter 8).
- Section 3 focuses on the post-design phases of evaluation, including conducting the evaluation (Chapter 9), analyzing data (Chapter 10), interpreting results (Chapter 11), and using evaluation findings (Chapter 12).
- Section 4 provides several case studies.

1.8 CHAPTER SUMMARY

Evaluation is a method used to determine the value or worth of something. In our case, that "something" is a program. A **program** is defined broadly in this text to include a group of activities ranging from a small intervention to a national or international policy. **Program evaluation** is evaluation used to determine the merit or worth of a program.

Evaluation has the following attributes, components, and purposes:

- **Systematic**: logical and organized; undertaken according to a plan
- **Operations**: evaluation processes involved with implementing the activities of a program
- **Outcomes**: results that occur during and after implementing a program
- **Standard**: the target used, implicitly or explicitly, to judge the merit or worth of a program
- **Improving programs and policies**: the purpose of evaluation; to create more effective and efficient programs and policies

Evaluation is a type of **research**, a systematic investigation in a field of study. **Embedded evaluation**, in particular, is an evaluation approach based on continuous improvement, in which program processes and practices are examined and refined to improve outcomes. A **stakeholder** is anyone who has an interest in or is involved with the operation or success of a program. Key stakeholder groups often include program staff, program participants, community members, and policymakers.

A **notice of funding opportunity** (NOFO) or **request for proposal** (RFP) is a solicitation for organizations to submit a proposal on how they would complete or address a specified project. Evaluation NOFOs and RFPs often ask for formative and summative evaluation. **Formative evaluation** is evaluation aimed at providing information to improve a program while it is in operation. Formative evaluation techniques include

- **Process evaluation:** formative evaluation aimed at understanding the operations of a program,
- **Implementation assessment:** formative evaluation that examines the degree to which a program is implemented with fidelity (according to plan),
- **Needs assessment:** formative evaluation that focuses on what services are needed and who needs them, and
- **Evaluability assessment:** formative evaluation used to determine if an evaluation is feasible and the role stakeholders might take in shaping the evaluation design.

Summative evaluation is evaluation aimed at providing information about effectiveness, to make decisions about whether to continue or discontinue a program. Summative evaluation techniques include

- **Outcome evaluation:** summative evaluation aimed at measuring how well a program met its stated goals,
- **Impact evaluation:** summative evaluation that measures both the intended and unintended outcomes of a program,
- **Cost-benefit/cost-effectiveness analysis:** summative evaluation that focuses on estimating the efficiency of a program regarding dollar costs saved (cost-benefit) or outcomes measured (cost-effectiveness),
- **Meta-analysis:** summative evaluation that integrates the effects of multiple studies to estimate the overall effect of a program, and
- **Meta-evaluation:** an evaluation of an evaluation.

Evaluators have many resources to guide their practice, including the American Evaluation Association's Guiding Principles for Evaluators, the Joint Committee's Program Evaluation Standards, and the American Evaluation Association's Evaluator Competencies. Finally, an **internal evaluator** is an evaluator employed by the organization that operates a program (but preferably not responsible for the program itself). An **external evaluator** refers to an evaluator who is employed from outside of the organization in which the program operates.

REFLECTION AND APPLICATION

1. Research the Drug Abuse Resistance Education (DARE) program. What have past evaluations found regarding the effectiveness of DARE?
2. Why do you think policymakers continue to support programs even after evidence suggests they are ineffective?
3. Explain why it might be considered unethical to not evaluate a program or policy.
4. Find a program or policy where data showed it was not working as intended or had unintended consequences. Was the program continued?
5. Describe embedded evaluation and how it is used.
6. Identify the steps of embedded evaluation.

KEY TERMS

See Glossary for all definitions.

AEA Evaluator Competencies
AEA Guiding Principles for Evaluators
Cost-benefit/cost-effectiveness analysis
Cost-utility analysis
Credibility
Embedded evaluation
Evaluability assessment
Evaluand

Evaluation
External evaluator
Formative evaluation
Impact evaluation
Implementation assessment
Improving programs and policies
Internal evaluator
Joint Committee’s Program Evaluation Standards
Meta-analysis
Meta-evaluation
Needs assessment
Notice of funding opportunity (NOFO)
Objectivity
Operations
Outcome evaluation
Outcomes
Process evaluation
Program evaluation
Program
Request for proposal (RFP)
Research
Stakeholder
Standard
Summative evaluation
Systematic
Value

2 HISTORY OF EVALUATION

Educating the mind without educating the heart is no education at all.

—**Aristotle**

LEARNING OBJECTIVES

2.1

- **a.** Discuss the historical context of evaluation.
- **b.** Describe how the discipline of evaluation has evolved over the last 200 years.
- **c.** Identify important contributors to the development of the field of evaluation.

2.2 Explain the history of research and evaluation with respect to ethics.

2.3 Identify current issues in evaluation.

2.1 THE EVOLUTION OF EVALUATION

While evaluation as a profession is new, evaluation activity began long ago, perhaps as early as Adam and Eve. As defined in Chapter 1, evaluation is a method used to determine the value or worth of something. It is a process humans use to make decisions. It is also an imperfect process. As humans, we evaluate with the information available to us, which is often incomplete and nearly always without a clear picture of implication and consequence. Adam and Eve made the decision to eat from the forbidden tree, evaluating the information that they had and obviously weighing one source more than another. Their information was conflicting and they did not foresee the consequences of their decision, but it was evaluative nonetheless. Some researchers look back further and place the roots of evaluation with evolutionary biology (Shadish et al., 1990). It is reasonable to consider that evaluation is at play when species mutate to adopt new characteristics as a survival adaptation, as with evolutionary developmental biology (Evo-Devo). Evo-Devo, no relation to the 1970s rock band Devo (which brings back many memories), is the study of when, how, and to what extent genes are turned on to maximize survivability through natural selection (Public Broadcasting Service, 2009). However, evaluation as an activity to improve processes, programs, and policies has more modest roots.

2.1.1 Before and During the 1800s

Beyond the evaluation associated with gene expression and the choices of Adam and Eve, evidence of evaluation has been documented as far back as 2200 BCE with the emperor of China's efforts to evaluate his staff every three years (Shadish et al., 1990; Wainer, 1987). About 1,000 years later, in 1115 BCE, the Chan dynasty began testing staff before they were hired; and over 2,000 years after that, in the late 1700s and early 1800s, France and then Britain adopted a similar assessment system for selecting civil servants (Wainer, 1987).

Also in Britain, in 1792, William Farish of Cambridge University is credited with creating the first system of grades (Hartmann, 2000; Soh, 2011). During the time, some universities in Britain had begun to base professor pay on the number of students they taught. An early entrepreneur of sorts, Farish developed a method to teach as many students as possible with the least amount of work, and thus make more money. His method was to assign quantitative grades to students. While some American universities, such as Yale, assigned categorical grades to students in the late 1700s, it was not until the early 1800s that quantitative grading schemes became popular in the United States (Schinske & Tanner, 2014).

In the early to mid-1800s, France and Britain began to look beyond evaluating people and toward evaluating programs and policies. One of the earliest examples of the evaluation of a social policy was in the 1830s by the French researcher André-Michel Guerry. Guerry studied how education relates to crime and concluded that education does not reduce crime (Cullen, 1975). This finding has been argued by statisticians both methodologically and with evidence (Weiss, 1998). Guerry also examined relationships between weather and mortality, as well as crime and suicide (Friendly, 2007). Further, in the 1840s, another French researcher, Jules Depuit, evaluated the usefulness of public works in France from an economic standpoint of supply and demand (Toulemonde & Rochaix, 1994). Also in the 1840s, Great Britain created commissions to focus on social problems. For instance, the Health of Towns Commission was formed to examine and improve conditions to decrease death rates in urban areas across England (British Broadcasting Corporation, 2014).

While there is evidence of the United States adopting systematic hiring practices during the late 1800s and group assessment to evaluate the intelligence of military recruits several decades later (Wainer, 1987), perhaps the first large-scale effort at evaluation in the United States was launched by Horace Mann in Boston. Mann, referred to as the "Father of Public Education," was dissatisfied with the Massachusetts education system (Cremin, 2018) and sought to create a free public school system that educated all citizens, regardless of race, religion, or income level (Baines, 2006; Gale Group, 2002). During the 1830s and 1840s, Mann advocated for education reform and pushed for objective assessment of student learning as a way to examine the effectiveness of Boston schools. The practice introduced by Mann in the mid-1800s, of using student test scores to evaluate educational programs, was the beginning of standardized testing in the United States and remains in use today (Hogan, 2007; Wozolek & Shafer, 2021). Due to the quantitative nature of early evaluative systems, many educators and lawmakers equated assessment and measurement to evaluation. That is, evaluation was narrowly seen as the quantitative assessment of outcomes.

2.1.2 Early to Middle 20th Century

Frederick Taylor, an inventor and engineer from Philadelphia, is known as the "Father of Scientific Management." His scientific management movement of the early 1900s was based on objective analysis of tasks and measurement of work outcomes to improve efficiency. Regardless of the many criticisms of scientific management (Locke, 1982), for example, that it did not recognize the more human side of management and employee performance, Taylor's methods of using data to foster change expanded the role of evaluation from mere description of assessment data to the use of those descriptive data for process improvement.

One of the earliest evaluations in social science is the Cambridge-Somerville Youth Study conducted in the 1930s. This study examined the effectiveness of welfare-type interventions, such as medical assistance, counseling, academic assistance, and community-based support, in preventing or reducing delinquency in at-risk boys. See "In the Real World" for more information on the Cambridge-Somerville Youth Study (Cabot, 1940; McCord, 1978, 2002, 2003; McCord & McCord, 1959; Welsh et al., 2023).

The first comprehensive, long-term evaluation in the field of education was conducted in Chicago between 1932 and 1940. The Eight-Year Study, spearheaded by Ralph Tyler, was an experiment across 30 secondary schools intended to test the effectiveness of different curricula (Alkin & King, 2016; Pinar, 2010). Tyler's work led to the exploration of national assessments in the United States, which resulted in the National Assessment of Educational Progress (NAEP). The development of the NAEP began in 1964, despite opposition from the American Association of School Administrators and the National Council of English Teachers, and was administered for the first time in 1969 (Vinovskis, 1998). NAEP, also referred to as the Nation's Report Card, is a government-mandated, common measure of achievement through which academic progress can be examined for the nation and by individual states, as well as for school districts that participate in the Trial Urban District Assessment (TUDA). TUDA currently includes over 25 urban school districts across the United States (National Center for Education Statistics, 2024). NAEP is administered by the National Center for Education Statistics (NCES) within the U.S. Department of Education's Institute for Education Sciences (IES); it has been used for over 50 years and currently tests across 10 content areas in Grades 4, 8, and 12 (NCES, 2019). Tyler, who was born in 1902, continued to contribute to the field through lecturing and consulting until his death in 1994. Because of his influence in the fields of assessment and evaluation, Ralph Tyler is referred to as the "Father of Evaluation" (Mukhongo, 2019).

IN THE REAL WORLD . . .

The Cambridge-Somerville Youth Study (CSYS) is one of the earliest evaluations funded by a private foundation, the Ella Lyman Cabot Foundation. The design of CSYS began in 1935 and took 4 years to complete. The purpose of CSYS was both to prevent juvenile delinquency among boys and to study the effectiveness of juvenile delinquency interventions.

Participants in the CSYS study were 650 school-aged boys. Boys were matched based on data from a 160-item code sheet including variables such as age, grade, physical health, intelligence, home life, and mental health. One boy in each of the 325 matched pairs was placed into either the Treatment (T) or Control (C) group; the matched boy was placed in the other group (i.e., if a boy was placed in T, his matched pair was placed in C or if a boy was placed in C, his matched pair was placed in T). Boys in the T group were assigned counselors and received specialized services from agencies in the Boston area. While there was some attrition, due to relocation or death, when possible, CSYS arranged for services to continue if a boy changed schools. The study was planned to last 10 years, with 2- to 3-year follow-ups during that time.

Data collected included variables related to personality development, community relationships, school progress, emotional maturity, medical problems, mental health, delinquency, and incarceration.

The theory behind CSYS was that interventions focused on character development, emotional security, social development, and related matters would decrease the likelihood of delinquency in boys during childhood and be preventive of later criminal activity.

Source: Cabot, P. S. deQ. (1940). A long-term study of children: The Cambridge-Somerville Youth Study. *Child Development, 11*(2), 143–151. https://doi.org/10.2307/1125845

While Tyler stands out as perhaps the most influential figure in early evaluation, the launching of Sputnik by the Soviet Union in 1957 helped propel the field of evaluation to where it is today. At the time, the United States thought itself the superpower of the world, yet the Soviet Union beat the Americans into space. Even with the United States following up with a successful launch of the Explorer 1 in 1958, the realization that the United States was not leading the space race called into question the effectiveness of the American education system in its ability to create top scientists. Sputnik led to the founding of the National Aeronautics and Space Administration (NASA) and initiated a new focus across America on technological and scientific discovery (Garber, 2007).

In addition to Sputnik, the postwar economy of the late 1940s through the 1960s was also an important factor in the development of the evaluation field. Along with the economic growth during this time came a greater call for social programs to bridge the gap between those who benefitted from current society and those who were suffering, living in poverty, and marginalized. In response to this call, some existing federal social programs were expanded and others created anew. As part of the Social Welfare History Project, Marx (2011) provides an overview of American social policy during the 1960s, including the following programs created and laws enacted during that time:

- The Juvenile Delinquency and Youth Offenses Control Act of 1961 funded programs aimed at reducing juvenile crime.
- In 1962, amendments to the Social Security Act created programs to aid families with dependent children.

- The Manpower Development and Training Act of 1962 created new job training programs.
- The Community Mental Health Centers Act of 1963 facilitated the creation of community mental health centers to provide preventive services.
- The Civil Rights Acts of 1964 and 1965 changed federal policy regarding the enforcement of sanctions for civil rights violations.
- In 1965, Medicare and Medicaid programs enabled senior citizens and those living in poverty to have access to health care.
- The Older Americans Act of 1965 formed a national network of organizations to serve the aging population with health and nutrition programs.
- The Elementary and Secondary Education Act of 1965 provided financial assistance to low-income schools.
- The Economic Opportunity Act of 1965 provided alternative training and job programs to youth.

Other social programs created during the 1960s included the federal Food Stamp Act, the Work Incentive program, the Work-Study program, and Head Start (Marx, 2011). It was during the second half of the 20th century, when social programs exploded and the focus on education expanded, that the field of evaluation was truly born.

2.1.3 Late 20th Century to Early 21st Century

Due to calls for educational reform following Sputnik and the proliferation of social programs in the 1960s, the need for critical examination of the effectiveness and impact of these reforms and programs became apparent. However, it also became apparent that professionals with the necessary evaluation skills were scarce. Further, the field lacked evaluative tools and methodologies with which to examine programs and policies. Thus, during the 1970s and born from a dearth of knowledge, the evaluation profession emerged. As the field developed, assessment remained a method to measure outcomes, but evaluation progressed beyond assessment to include additional methods and approaches.

During the 1970s, professionals from many domains contributed to the development of evaluation as a field in its own right. Psychologists, including Ralph Tyler, Lee Cronbach, and Donald Campbell, brought quantitative methods to evaluation. Sociologists, such as Michael Quinn Patton and Carol Weiss, developed qualitative and theory-based approaches to evaluation. Other early evaluators from the realms of philosophy, communications research, educational psychology, and statistics helped shape the wealth of evaluation tools and approaches we have today. See Table 2.1 for a list of important contributors to the field of evaluation, as well as where they were employed (if applicable) in 2024, where they studied, and their field of study. This table is not meant to be exhaustive and surely there are important contributors to the field that are not included, but it serves as a starting point for understanding the convergence of many disciplines to shape evaluation as a profession. The list includes evaluators who helped lay the foundation for the field of evaluation and those who continue to shape the field today. The particular contributions of individuals to the field of evaluation, as well as a discussion of evaluation approaches, are addressed in more detail in Chapter 4.

The first university courses on evaluation were also developed in the 1970s. Some of the pioneering institutions were Stanford University, Western Michigan University, and the University of Illinois (Hogan, 2007). While funding for program evaluation at the federal level was cut in the 1980s, the field continued to develop and expand. In the mid-1980s, the **American Evaluation Association (AEA)** was created when two smaller associations, the Evaluation Research Society (ERS) and the Evaluation Network (ENet), merged (Kingsbury, 1986). The AEA is an international professional association of evaluators focused on sharing knowledge of evaluation approaches and methods. The group hosted its first annual conference in 1986. In 2024, the AEA had approximately 5,000 members across all 50 U.S. states and more than 80 countries (see www.eval.org)

During the 1990s, the U.S. government increased funding for and amplified the focus on program and policy evaluation of federal initiatives. States and local organizations began to look to evaluation as a way to improve their programming. The states of Massachusetts and South Carolina as well as the city of Chicago included evaluation in their human services programs (Weiss, 1998). Many evaluation texts and journal articles were published in the 1990s, adding to the wealth of resources available to evaluation professionals. Due to the diverse backgrounds of early evaluators (see Table 2.1), the approaches and methods of evaluation varied considerably, which sparked debate over the merit and worth of different approaches. These debates continue today, but it is through this debate and dialogue that evaluators have formed a community of professional learning where divergent thinking can be discussed, critiqued, advanced, and, most important, respected. Evaluation approaches, including their strengths and critiques, are reviewed in Chapter 4.

TABLE 2.1 ■ Contributors to the Field of Evaluation

Evaluator	Employment (if applicable, as of 2024)	Degree Received From/ Date	Degree/Field of Study
Marvin Alkin	UCLA	Stanford University (1964)	EdD, Education
Thomas Archibald	Virginia Tech	Cornell University (2013)	PhD, Adult and Extension Education
Michael Bamberger	Independent Consultant	London School of Economics (1965)	PhD, Sociology
Donald T. Campbell	*Deceased 1996*	University of California, Berkeley (1947)	PhD, Psychology
Eleanor Chelimsky	*Deceased 2022*	University of Paris[1]	Diplome Superieur, Music
Huey T. Chen	Mercer University	University of Massachusetts—Amherst[1]	PhD, Sociology

Evaluator	Employment (if applicable, as of 2024)	Degree Received From/ Date	Degree/Field of Study
Leslie J. Cooksy	LJC Consulting	Cornell University (1989)	PhD, Program Evaluation and Public Policy
Thomas D. Cook	Northwestern University	Stanford University (1967)	PhD, Communications Research
J. Bradley Cousins	University of Ottawa	University of Toronto (1988)	PhD, Educational Measurement and Evaluation
Lee J. Cronbach	*Deceased 2001*	University of Chicago (1940)	PhD, Educational Psychology
E. Jane Davidson	Real Evaluation	Claremont Graduate University (2001)	PhD, Psychology
Jara Dean-Coffey	jdcPARTNERSHIPS; Equitable Evaluation Initiative	University of California, Berkeley (1997)	MPH, Community Health and Gerontology
Stewart I. Donaldson	Claremont Graduate University	Claremont Graduate University (1991)	PhD, Psychology
Stephanie Evergreen	Evergreen Data	Western Michigan University (2011)	PhD, Evaluation
David M. Fetterman	Fetterman & Associates	Stanford University (1981)	PhD, Educational and Medical Anthropology
Jennifer C. Greene	University of Illinois	Stanford University (1976)	PhD, Educational Psychology
Gary T. Henry	University of Delaware	University of Wisconsin–Milwaukee (1982)	PhD, Political Science
Stafford Hood	*Deceased 2023*	University of Illinois at Urbana-Champaign (1984)	PhD, Program Evaluation
Rodney Hopson	American University	University of Virginia (1997)	PhD, Educational Evaluation
Ernest R. House	University of Colorado Boulder	University of Illinois at Urbana-Champaign (1968)	EdD, Education
Jean A. King	University of Minnesota	Cornell University (1979)	PhD, Curriculum and Instruction
Mark W. Lipsey	Vanderbilt University	Johns Hopkins University (1972)	PhD, Psychology

(*Continued*)

TABLE 2.1 ■ Contributors to the Field of Evaluation (*Continued*)

Evaluator	Employment (if applicable, as of 2024)	Degree Received From/ Date	Degree/Field of Study
Mel M. Mark	Pennsylvania State University	Northwestern University (1979)	PhD, Psychology
Donna Mertens	Galluadet University	University of Kentucky (1977)	PhD, Educational Psychology
Bianca Montrosse-Moorhead	University of Connecticut	Claremont Graduate University (2009)	PhD, Psychology (Evaluation and Research Methods)
Frederick Mosteller	*Deceased 2006*	Princeton University (1946)	PhD, Mathematics
Hallie Preskill	Hallie Preskill Consulting	University of Illinois at Urbana-Champaign (1984)	PhD, Program Evaluation
Michael Quinn Patton	Consultant, Utilization-Focused Evaluation	University of Wisconsin–Madison (1973)	PhD, Sociology
Peter H. Rossi	*Deceased 2006*	Columbia University (1951)	PhD, Sociology
Michael J. Scriven	*Deceased 2023*	Oxford University (1956)	PhD, Philosophy
William R. Shadish	*Deceased 2016*	Purdue University (1978)	PhD, Clinical Psychology
Robert E. Stake	University of Illinois at Urbana-Champaign	Princeton University (1958)	PhD, Psychology
Daniel L. Stufflebeam	*Deceased 2017*	Purdue University (1964)	PhD, Statistics and Measurement
William M. K. Trochim	Cornell University	Northwestern University (1980)	PhD, Psychology
Ralph W. Tyler	*Deceased 1994*	University of Chicago (1927)	PhD, Educational Psychology
Carol H. Weiss	*Deceased 2013*	Columbia University (1977)	PhD, Sociology
Joseph S. Wholey	*Deceased 2023*	Harvard University[1]	PhD, Philosophy

[1] Unable to locate year of degree.

2.1.4 Hogan's Framework

While the historical evolution of evaluation can be explored by century, it can also be examined at a finer level. Hogan's (2007) framework of evaluation development provides a rich conceptualization of how and when the field developed. He divides the progression of program evaluation into seven time periods, beginning in the late 1700s:

1. Age of Reform (1792–1900s)
2. Age of Efficiency and Testing (1900–1930)
3. Tylerian Age (1930–1945)
4. Age of Innocence (1946–1957)
5. Age of Development (1958–1972)
6. Age of Professionalization (1973–1983)
7. Age of Expansion and Integration (1983–)

Hogan describes the Age of Reform as the time when the first recorded evaluation took place. As mentioned previously, higher education institutions in England and the United States began to use quantitative methods to evaluate students in the late 1700s, partially in an effort to increase income by teaching a greater number of students. Similar measures were used to assess the performance of civil servants and in hiring practices for military recruits. Beyond evaluating people, measures of student learning were used to examine the performance of educational programs and systems. After the turn of the 20th century, during what Hogan calls the Age of Efficiency and Testing, Taylor's scientific management further facilitated the movement toward objective measurement and assessment as a form of evaluation. Ralph Tyler, the "Father of Evaluation," has his own era, the Tylerian Age. It was during this time that objectives were used as a foundation for evaluation. Tyler's work on national assessments across multiple content areas is still evident today.

The Age of Innocence, during the mid-1900s, is aptly labeled by Hogan, because it refers to the time when many programs were created and investments made in the United States, without thought to whether they were worth the time and money allocated to them. That is, the postwar economy spurred both intense need and rapid growth across many sectors, in what some might call an irresponsible rollout of actions without regard to long-term consequences. Hogan's label of "innocence" is nicer than mine of "irresponsibility." The Age of Development, immediately following the launch of Sputnik, propelled the United States further into growth mode; however, conversations arose regarding accountability and questions were asked about effectiveness of the many investments made in the preceding decades.

My favorite of Hogan's ages, the Age of Professionalization, is the time during which the evaluation field took on its modern contours. Due to the clear need to examine the effectiveness of government spending and associated programs, as well as the call for evaluation from private foundations and organizations, researchers across many fields converged and joined forces to develop the emerging field of evaluation. Professional organizations were formed, evaluation methodologies generated, methods of dissemination (such as journals) created, and university programs focused on producing evaluation professionals developed. Finally, Hogan's last stage was the Age of Expansion and Integration, for which no end date was offered. During this time, evaluation as a field extended across disciplines and methods. The sweeping cuts in social programs of the 1980s gave way to increased globalization, environmentalism, and technological advances of the 1990s. The shrinking of resources for evaluation during the 1980s eased and funding (as well as expectation) for rigorous evaluation multiplied. A focus on accountability for government-funded programs spurred the need for trained evaluators. The two primary evaluation associations that were created in the mid-1970s and merged in the mid-1980s, the Evaluation Network and the Evaluation Research Society, became more integrated as university-based and government evaluators worked together to shape the new AEA.

The field of evaluation has continued to grow since Hogan's Framework was published in 2007. The AEA is well established with a diverse group of evaluators who collaborate collegially yet organize themselves across topical interest groups (TIGs). As such, I humbly claim that we entered an eighth age in 2010 with Leslie Cooksy's keynote address on evaluation quality at the AEA Annual Conference and followed by her manuscript with Mel Mark on the various influences of evaluation quality (Cooksy & Mark, 2012).

8. Age of Acuity (2010–present)

This time period, which I refer to as the Age of Acuity, continues to this day. With evaluation designs and methods established and expanding, the focus on data-driven decision making more widespread, and evaluators organized and supported by infrastructure for collaboration, evaluation advances in the 21st century underscore and illuminate the importance of values, vision, equity, and multiculturalism. Factors that are important to consider in evaluation are receiving attention and as a result, evaluators are becoming more aware, responsive, and capable of addressing context and culture, as well as confronting issues of equity, power, and inclusion in their approach to evaluation. Yet evaluation is still a relatively young field, and there remains countless opportunities for discovery and growth (and contribution), as evidenced by the many individuals in Table 2.1 who are active evaluators and continue to shape the field.

IN THE REAL WORLD . . .

The Cambridge-Somerville Youth Study (CSYS) was described earlier in the chapter (see "In the Real World"). About 325 boys received counseling, family guidance, academic assistance, medical assistance, and other community-based services over a 5-year period between 1939 and 1944.

Findings 3 years post-program, as analyzed by Powers and Witmer (1951), revealed no significant impact on criminal behavior. That is, boys who participated in the program fared no better (or worse) than those who did not participate in the program regarding criminal offenses.

Findings after 15 years as analyzed by McCord and McCord (1959) also showed little evidence that the program had reduced criminal behavior. They concluded that the intervention provided to treatment boys through CSYS was ineffective at crime prevention. However, from subsequent analyses, they did find that boys who began treatment earlier (before 10 years of age) and those who had more interaction with their counselor/social worker (weekly visits) had less criminal behavior than boys who started treatment later and had less frequent contact with their social worker.

Findings after 30 years, as analyzed by McCord (1978), not only showed no evidence that the CSYS program reduced crime, but also revealed that boys who participated in the program had poorer later life outcomes than boys in the control group. As adults, boys who participated in the program were more likely to commit a second crime, more likely to show signs of alcoholism and serious mental illness, more likely to have a stress-related disease, and more often reported dissatisfaction with their job. McCord hypothesizes that interaction with a counselor during childhood may foster dependency on outside services and create expectations of success that were not realized. She concludes that social work interventions, such as CSYS, actually increase risk of poor later life outcomes for the youth they are designed to help.

Criticisms of McCord's analyses and subsequent conclusion regarding childhood counseling came from many fronts. Researchers believed her study lacked rigor and neglected to use more sophisticated analyses that might have been more informative (Vosburgh & Alexander, 1980). Conclusions based on treatment versus control boys have also been criticized because the control group was not a "no treatment" group, but more likely a group of boys who received other services that were not documented in the study.

Other hypotheses about why the treatment boys had poorer long-term outcomes include McCord's (2003) peer deviancy theory. She believed the CSYS intervention component in which treatment boys were brought together at a camp fostered social connections that may have allowed deviant youth to bond and reinforce deviant behavior (McCord, 2002). McCord also found that parenting practices and behavior had a greater impact on outcomes than family structure. In particular, boys whose fathers committed serious criminal offenses were more likely to commit crimes themselves; however, this relationship did not hold for boys whose fathers committed criminal offenses but were absent from the family structure. Other researchers point to subsequent research on protective factors, social influences, and institutional influences (Welsh et al., 2017; Welsh et al., 2023) as influences on later life criminality.

A fourth follow-up is currently underway. Thus far, findings after 70 years, as analyzed by Welsh and colleagues (2023), have shown no effects on mortality. However, it was found that participants who had been criminal offenders throughout their life had greater mortality, especially from unnatural causes, than participants who had criminal offenses only in adolescence and those who had no criminal offenses.

Sources: McCord, J. (1978). A thirty-year follow-up of treatment effects. *American Psychologist, 33*(3), 284–289. https://doi.org/10.1037/0003-066X.33.3.284; McCord, J. (2002). Counterproductive juvenile justice. *Australian and New Zealand Journal of Criminology, 35*(2), 230–237; McCord, J. (2003). Cures that harm: Unanticipated outcomes of crime prevention programs. *Annals of the American Academy of Political and Social Science, 587*(1), 16–30. https://doi.org/10.1177/0002716202250781; McCord, J., & McCord, W. (1959). A follow-up report on the Cambridge-Somerville Youth Study. *Annals of the American Academy of Political and Social Science, 322*(1), 89–96. https://doi.org/10.1177/000271625932200112; Powers, E., & Witmer, H. L. (1951). An experiment in the prevention of delinquency: The Cambridge-Somerville Youth Study. Columbia University Press; Vosburgh, W. S., & Alexander, L. B. (1980). Long-term follow-up as program evaluation: Lessons from McCord's 30-year follow-up of the Cambridge-Somerville Youth Study. *American Journal of Orthopsychiatry, 50*(1), 109–124; Welsh, B. C., Zane, S. N., & Rocque, M. (2017). Delinquency prevention for individual change: Richard Clarke Cabot and the making of the Cambridge-Somerville Youth Study. *Journal of Criminal Justice, 52*(C), 79–89. https://doi.org/10.1016/j.jcrimjus.2017.08.006; Welsh, B. C., Zane, S. N., Yohros, A., & Paterson, H. (2023). Cohort profile: The Cambridge-Somerville Youth Study (CSYS). *CrimRxiv.* https://doi.org/10.21428/cb6ab371.9c13676d

QUICK CHECK

1. Who is considered the "Father of Evaluation" and why?
2. What event occurred in the 1950s that helped jump-start the field of evaluation? What occurred during the 1960s that further brought to bear the need for qualified evaluators?
3. During what time frame was evaluation recognized as a profession? What fields did the early contributors to evaluation approaches come from?
4. What is the primary professional association for evaluators?

2.2 THE HISTORY OF ETHICS IN RESEARCH AND EVALUATION

Much of what we know about modern medicine, human behavior, and effective practices is due to research. As stated in Chapter 1, research and evaluation are necessary to ensure that the programs and policies, as well as treatments and interventions, we use work for the people they are designed to help. In a sense, it is an ethical obligation of researchers, whether they are basic scientists or program evaluators, to examine whether resources are being spent wisely on methods that are effective. As researchers and evaluators, we seek to increase and share knowledge to the betterment of human beings. However, what trumps this knowledge-generation process is that no harm is done along the way. Unfortunately, in the United States and around the world, there have been numerous experiments performed on humans, perhaps aimed at the greater good, but without regard for the human beings that were exploited. In some cases, the harm to humans has been deliberate and callous, where individuals were dehumanized and seen solely as test subjects. In other cases, researchers may have been more neglectful than outright malicious, but the end result was the same: harm to people. Because of the sometimes horrific and always troubling abuses of humans in the name of research, guidelines and protections for humans involved in research have been established in the last 50 years. Researchers and evaluators alike are bound by these ethical guidelines. Guidelines and protections for humans involved in research are discussed in Chapter 4; this chapter examines the history of unethical treatment of humans during research that led to the need for ethical guidelines and oversight.

2.2.1 Human Experimentation Outside of the United States

Nazi Germany Experiments. Experiments conducted by Nazis during World War II were inarguably the worst abuses of humans in history. Nazis experimented on millions of individuals, including men, women, and children. Experiments were conducted on humans without their consent and without regard to pain and suffering. Individuals were exposed to freezing temperatures, poison, tuberculosis, sterilization, joint transplants, toxic gas, and infections (Tyson, 2000).

Japanese Unit 731. During and after World War II, it is reported that Japan experimented on potentially hundreds of thousands of men, women, and children using chemical and biological warfare.

These experiments, called Japanese Unit 731, also included vivisection, limb amputations, and freezing experiments similar to those performed in Nazi Germany (Kristoff, 1995).

Soviet Chamber. Prior to World War II and operating until at least the 1950s, the Soviet Union had a secret laboratory it called the Chamber. The Chamber was used to experiment with deadly poisons on humans (Central Intelligence Agency, 1993).

Aversion Project. During the 1970s and 1980s, South Africa conducted experiments to convert homosexuals to heterosexuals. Lesbian and gay soldiers were forced to undergo hormone treatments and even chemical castration. In addition, gender reassignment surgery was performed, without consent, on nearly a thousand men and women (Kaplan, 2004). This massive experiment on homosexuals is commonly referred to as the Aversion Project.

2.2.2 Human Experimentation Within and By the United States

Tuskegee Syphilis Experiments. For 40 years beginning in 1932, the U.S. Public Health Service and the Tuskegee Institute in Tuskegee, Alabama, experimented on African American farmers who lived in poverty to learn about the progression of and treatments for syphilis. Six hundred men, about 400 with syphilis and 200 without, were given free medical care in return for their participation in the study. Even when penicillin became recognized as an effective treatment for syphilis in 1947, study participants were not offered this treatment. In 1997, President Bill Clinton apologized to the eight surviving participants of the Tuskegee experiments (Physicians Committee for Responsible Medicine, 2019).

Monster Study. To test his theory that the diagnosis of stuttering can itself cause stuttering, a University of Iowa researcher, Wendell Johnson, conducted a study in 1939 with children at an orphanage. Orphaned children with normal speech patterns were told they had poor speech, including a stutter. These children, who did not have speech problems prior to the study, developed stutters and suffered negative psychological and behavioral effects (Silverman, 1998).

U.S. Radiation Experiments. From the mid-1940s until the 1980s, the U.S. government conducted a research program focused on the effects of radiation on humans. Hundreds of experiments were sponsored across the United States; subjects included the elderly, prisoners, pregnant women, and terminally ill patients. These experiments were conducted at multiple sites, and in many cases subjects received radiation doses up to 98 times greater than what was known at the time to be tolerable (Faden, 1996; Knight-Ridder, 1994; U.S. Department of Energy, 1995; U.S. House of Representatives, 1986).

Guatemala Syphilis Experiments. In 2010, while examining documents from the Tuskegee syphilis study, a researcher discovered that a similar experiment was performed by the U.S. government between 1946 and 1948 in Guatemala. The experiment was called the U.S. Public Health Service Sexually Transmitted Disease Inoculation Study. Over 1,300 people were intentionally infected with venereal diseases, including syphilis and gonorrhea, to examine how effective penicillin was in treating the diseases. Only a portion of the subjects were administered penicillin and over 80 individuals died from participation in the study

(Fox, 2010; Resnick, 2019; Rodriquez & Garcia, 2013). On October 1, 2010, President Barack Obama apologized to the Guatemalan people on behalf of the U.S. government.

Project MK-Ultra. The Central Intelligence Agency (CIA) conducted mind control experiments called MK-Ultra, beginning during the Cold War in the 1950s and continuing through the 1960s. Participants were exposed to hallucinogenic drugs such as LSD, hypnosis, radiation, toxins, chemicals, electroshock, and lobotomy as part of the CIA's research into behavior modification. While some subjects agreed to participate, many were coerced or did not even know they were involved in an experiment. Subjects included mentally impaired boys, American soldiers, mental hospital patients, and prisoners. Due to the records being destroyed by the CIA in 1973, the government was unable to identify all who participated (Budiansky & Goode, 1994; Nofil, 2019).

Holmesburg Prison Experiments. Beginning in the early 1950s, a researcher from the University of Pennsylvania School of Medicine, Albert Kligman, paid prisoners at Holmesburg Prison a small fee to perform a variety of experiments on them. Prisoners were infected with ringworm, herpes, and staphylococcus; were exposed to toxic drugs and chemicals; participated in commercial testing for products such as detergents and dyes; and were used by pharmaceutical companies to test drugs, including tranquilizers and antibiotics. Inmates suffered many side effects including hallucinations, skin lesions, scars, memory loss, and cognitive impairment. Even with ethical codes being established due to the atrocious experiments by the Nazis, these experiments continued until they were finally stopped in the mid-1970s (Hornblum, 1998).

Milgram Obedience Experiments. In an effort to understand why German military personnel followed orders and took part in the horrendous Nazi experiments during World War II, in 1961 Stanley Milgram, a psychologist at Yale University, undertook a series of "obedience" experiments. The Milgram experiments studied how far people would go to obey authority. Study participants were told to shock a "learner" for incorrect answers; however, the study participants did not know the learner was not a real person, but rather a recording. After each shock, the participant was instructed to increase the voltage of the next shock, despite the learner's call for them to stop. If the study participant hesitated, the authority figure prodded the participant to continue with the experiment. Nearly two thirds (65%) of participants obeyed the authority figure to the point of maximum shock. While Milgram debriefed participants after the experiment about the deception and the true purpose of the experiment, these experiments have been highly criticized and are deemed by most to be ethically questionable and by many to be unethical (Miller et al., 1995).

Tearoom Trade Study. In 1970, Laud Humphreys conducted a study to understand impersonal sex in public restrooms, called "tearooms." Humphreys, a doctoral student at Washington University in St. Louis, Missouri, documented the experience through field notes while serving as the lookout, or "watchqueen," during the casual sexual encounter (Humphreys, 1975). Subjects did not know Humphreys was a researcher, that he was taking field notes, or that he followed the men to their cars to record their license plate numbers. Using public records, he located their home addresses and visited their homes a year later under the guise of a mental health interviewer. Of the 134 men for whom he had located home addresses, 50 agreed to an interview. While Humphrey's findings serve to dispel inaccurate stereotypes, his tactics have received much criticism because individuals did not consent to participate in his research (Nardi, 1995).

Stanford Prison Experiments. In 1971, Philip Zimbardo from Stanford University conducted an experiment to study people's psychological reactions to being held captive. Participants were male college students who volunteered, in exchange for financial compensation, to be in a psychological study simulating a prison. The study was designed to last 2 weeks, but was terminated after 6 days due to abusive conditions and psychological distress. While most agree that the experiment violated ethical standards, there is continuing discussion about why participants who were assigned to be prison guards so quickly took on inhumane, power-hungry behaviors, and why the subjects who were assigned to be prisoners accepted this treatment. Some believe it was due to the power inherent in the simulated situation, while others believe personal disposition was a factor. Regardless, the student volunteers suffered psychologically as a result of their participation in this experiment (Carnahan & McFarland, 2007).

2.2.3 Human Experimentation Today

The experiments described above are certainly not all the unethical studies conducted in the United States and beyond over the past century; however, they are some of the most notorious. They shaped a history of ethical violations in research on humans that led to explicit protections of humans and clear guidelines for researchers. These guidelines apply to all researchers, including evaluation researchers. The history of legislation related to human-subject protections and ethical conduct of researchers are reviewed in Chapter 4.

Even with all that has been done to humans in the name of research, and our ability to retrospectively identify ethical violations, have we really learned? There are always new areas of research that may not be explicitly addressed in current ethical standards, for example, in gene research and modification. As researchers, it is important that we always be reflective and deliberate in our actions. In 2015, the National Institutes of Health (NIH) declared that it would not fund research that employs gene editing of human embryos (NIH, 2015). Three years later, the director of the NIH, Francis Collins, released a statement expressing concern over human-genome editing, after a Chinese researcher had just released news that the first gene-edited twin babies had been born in China (NIH, 2018). The NIH reaffirmed its position of not supporting gene editing of human embryos. In 2019, Collins and the NIH called for an international moratorium on human-gene editing and the alteration of DNA before implantation. Other countries have joined this moratorium, but it is not yet a policy supported by all nations (NIH, 2019).

I had the pleasure of hearing Frances Collins speak a few years ago and was struck by his strong ethical convictions. He clearly supports science and research, but not without a firm grasp of the implications that scientific advances may have on the future. He is not asking researchers to never explore this area, but he is asking researchers worldwide to have a discussion about how laboratory-modified genes might affect humans. Science may allow us to create a genetically modified baby, but that baby will grow up. Until we have a clear understanding of how our interfering with the creation of human life might affect that human life (and all human life), we should proceed with measured steps and the utmost caution. Perhaps if some of the earlier researchers had taken a step back before embarking on an experiment, and really weighed the potential intended and unintended consequences, we would not have such a checkered past of ethical shortcomings. We should conduct research, not because we can, but because it is right.

2.3 COMMON THREADS AND CURRENT ISSUES IN EVALUATION

There are many relevant and timely issues in evaluation that are covered throughout the text. In this section, we discuss some common issues in evaluation as well as infrastructure supports that aid in addressing these issues.

2.3.1 Shadish's Common Threads

In his editorial "The Common Threads in Program Evaluation," William Shadish (2006) identified five concerns that appear throughout the program evaluation literature:

Concern 1: How do evaluators construct knowledge about programs?

Concern 2: How do evaluators place value on evaluation results?

Concern 3: How do programs change and how can evaluation be used to influence that change?

Concern 4: How do evaluators use evaluation results to influence policy making?

Concern 5: How can evaluators organize their practice to address concerns 1–4?

These concerns, or "common threads," have arisen from and helped shape the field of evaluation, and these concerns still permeate every meeting of the AEA. Concern 1 speaks to how we conduct evaluation, including what we can and cannot measure and the approaches and designs we use to understand a program's operation and impact. Concern 2 relates to the theoretical frameworks and practical methods that help evaluators make sense of evaluation results and value results such that they can inform recommendations. Concern 3 is one of the primary differences between basic research and program evaluation. Program evaluation is intended for practical use and application such that program activities can be improved. The usefulness and use of evaluation findings are necessary for change to occur. Concern 4 is similar to Concern 3, but relates to leveraging evaluation results to influence the policy process. To leverage findings, evaluators need to identify facilitators and barriers to use by policymakers and work to share results in such a way that capitalizes on facilitators, overcomes barriers, and ultimately advances the use of data in the policy-making process. Finally, Concern 5 is about organizing our practice as evaluators, to balance the methods used in conducting an evaluation, the way in which results are communicated, how these results are used for program improvement, and the extent to which findings can influence the policy process.

2.3.2 Emerging Issues in Evaluation

In her blog "7 Ways Program Evaluation Has Changed in 15 Years," Payal Martin (2023) reflects on how evaluation is different today than it was 15 years ago. Three items on her list are related to themes already mentioned in Chapters 1 and 2: (1) more demand for timely and ongoing data to support continuous program improvement, (2) increased mindfulness of the ethical obligations of evaluators, and (3) an intensifying call for identifying evidence-based strategies and providing metrics related cost-effectiveness and cost-benefit. The latter issue has implications for how we design our evaluations. Rigorous methods, such as randomized controlled experiments and regression discontinuity designs, are necessary for making causal assertions, as well as for meeting the standards required for a program to be identified as evidence-based.

Two areas on her list relate to evaluation approaches and methods: (4) an increased focus on engaging stakeholders in the evaluation process and (5) the use of mixed methods, including incorporating storytelling as a way to clearly communicate evaluation findings. The last two topics are perhaps the most emergent and have great consequence to the field of evaluation: (6) an emphasis on evaluation approaches that are culturally responsive, equitable, and inclusive; and (7) technological innovations such as artificial intelligence and machine learning. These two issues also intersect, in that careful attention must be paid to how the use of artificial intelligence (AI) in evaluation attends to diversity, equity, inclusion, and accessibility (DEIA). While culturally responsive and equitable evaluation (CREE) is discussed in Chapter 4, the approach is also briefly mentioned in this section as it relates to AI.

In his *AEA365* blog, Fetterman (2023) discussed how AI can be used to advance evaluation capacity. He included examples such as developing logic models and creating evaluation questions. While AI can be used to aid in evaluation, evaluation can also be used to inform the national conversation around AI use. The National Artificial Intelligence Advisory Committee (NAIAC) was created in response to Biden's Presidential directive on the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (Exec. Order No. 14110, 2023). The NAIAC is comprised of 26 experts in AI. They are tasked with providing recommendations on how AI is developed, governed, and equitably deployed. A key component to this is developing metrics to evaluate AI technologies.

The *New Directions in Evaluation* journal dedicated an issue to AI (Montrosse-Moorhead & Mason, 2023). While all papers in the issue provide valuable information on and insight into the use of AI in evaluation, in the final paper, Montrosse-Moorhead (2023) defines eight criteria for evaluating how AI is used in evaluation. Two of the criteria relate to equity in process and equity in findings, that is, whether AI advances equity and addresses inequities, including those related to race, gender, ethnicity, sexual orientation, and disability. Three criteria attend to trustworthiness and validity, as well as understandability, of findings produced by AI. The final three criteria pertain to how AI is used in evaluation design and implementation, whether the process of using AI is efficient, and the effectiveness of findings generated by AI. The use of AI for evaluation and the evaluation of AI technologies are important emerging issues in evaluation.

2.3.3 Resource Sharing and Dissemination

The Cochrane Collaboration is an international organization that provides synthesized research evidence around topics in health care. The Collaboration was created in 1993 in the United Kingdom as a way to facilitate the sharing and promote the use of evidence-based practices and interventions in health care decision making. Cochrane can be accessed at https://www.cochrane.org/.

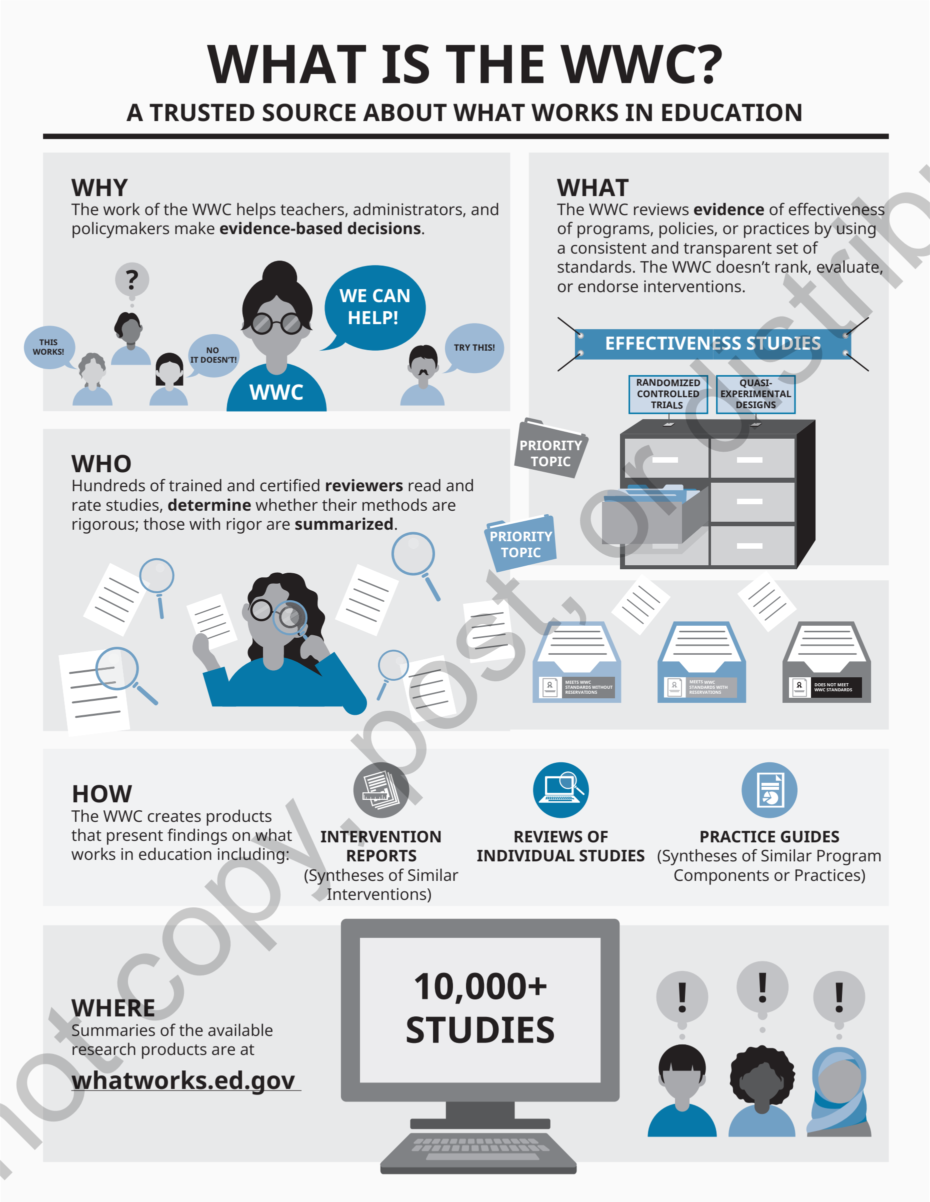
The Campbell Collaboration was created in 2000 based on the Cochrane Collaboration model. It is named after Donald Campbell, a psychologist who helped shape the field of scientific inquiry in the social sciences. Just as the Cochrane Collaboration focuses on systematic reviews in the medical and health fields, the Campbell Collaboration focuses on systematic reviews of social and behavioral interventions and programs. The Campbell Collaboration can be accessed at https://campbellcollaboration.org/.

The What Works Clearinghouse (WWC) is a resource provided by the U.S. Department of Education's Institute of Education Sciences (IES). It was created in 2002 with the involvement of some of the same researchers who helped start the Campbell Collaboration. The WWC includes evidence-based practices and programs in many education-related areas, including literacy, STEM, social-emotional learning and behavior, school leadership, college readiness, and career and technical education. The Clearinghouse reviews individual studies according to a standardized protocol and assigns one of three *WWC ratings*: (1) meets WWC standards without reservations, (2) meets WWC standards with reservations, or (3) does not meet WWC standards. Findings from WWC studies where the design meets standards without or with reservations are then examined for evidence of effectiveness. A study may be determined to have no statistically significant findings or statistically significant findings in one or more areas. Statistically significant findings are categorized by level of effectiveness; *effectiveness levels* are positive or potentially positive findings, mixed findings, negative or potentially negative findings, or no discernable findings. The WWC can be accessed at https://ies.ed.gov/ncee/wwc/. See Figures 2.1 and 2.2 for information on the WWC and how it rates evaluation studies.

In addition to reviewing individual studies, the WWC creates intervention reports through a rigorous review process based on ratings across multiple studies of a program or strategy. These intervention reports include how many studies were reviewed, how many of the revised studies meet WWC standards for inclusion, the outcome areas included in the studies, a systematically assigned *effectiveness level* by outcome area, and an *evidence tier* based on strength of evidence. Effectiveness levels were discussed in the preceding paragraph; the assigned rating is based on significance, magnitude, and consistency of findings. See Figure 2.3 for a description of the effectiveness levels used in intervention reports.

Evidence tiers were added to the review protocol due to federal legislation. These tiers are defined within the Every Student Succeeds Act (ESSA), the national education law passed on 2015 that replaced the No Child Left Behind legislation. ESSA strongly advocates for the use of evidence-based practices. The strength of evidence tier rating is determined by systematically examining a study based on five factors: (1) the design of a study, (2) sample size and setting of a study, (3) the study results, (4) findings from other studies for the same or similar program, and (5) how well the sample used in and setting of a study matches

FIGURE 2.1 ■ What Works Clearinghouse



WHAT IS THE WWC?

A TRUSTED SOURCE ABOUT WHAT WORKS IN EDUCATION

WHY

The work of the WWC helps teachers, administrators, and policymakers make **evidence-based decisions**.

?
THIS WORKS!
NO IT DOESN'T!
WE CAN HELP!
WWC
TRY THIS!

WHAT

The WWC reviews **evidence** of effectiveness of programs, policies, or practices by using a consistent and transparent set of standards. The WWC doesn't rank, evaluate, or endorse interventions.

EFFECTIVENESS STUDIES
RANDOMIZED CONTROLLED TRIALS
QUASI-EXPERIMENTAL DESIGNS
PRIORITY TOPIC
PRIORITY TOPIC
MEETS WWC STANDARDS WITHOUT RESERVATIONS
MEETS WWC STANDARDS WITH RESERVATIONS
DOES NOT MEET WWC STANDARDS

WHO

Hundreds of trained and certified **reviewers** read and rate studies, **determine** whether their methods are rigorous; those with rigor are **summarized**.

HOW

The WWC creates products that present findings on what works in education including:

INTERVENTION REPORTS (Syntheses of Similar Interventions)

REVIEWS OF INDIVIDUAL STUDIES

PRACTICE GUIDES (Syntheses of Similar Program Components or Practices)

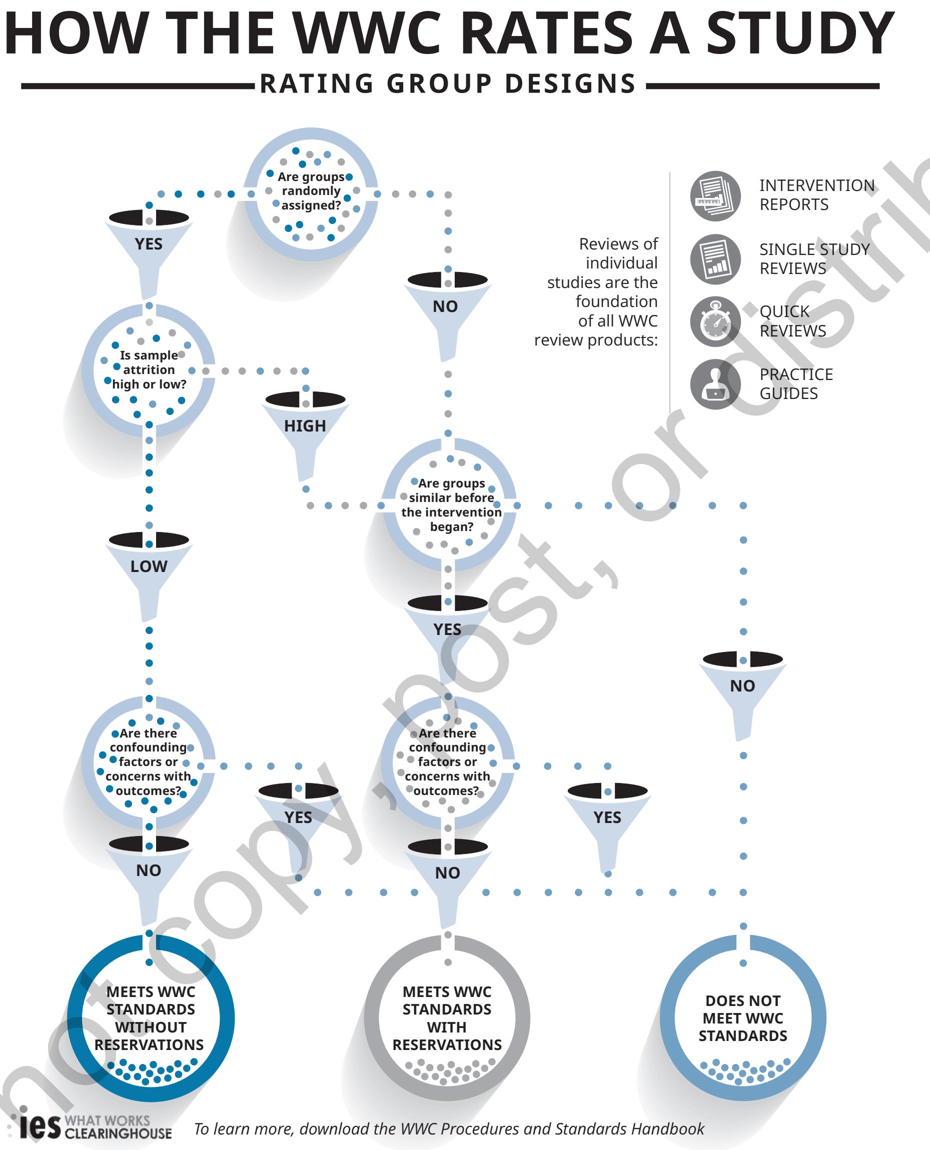
WHERE

Summaries of the available research products are at

whatworks.ed.gov

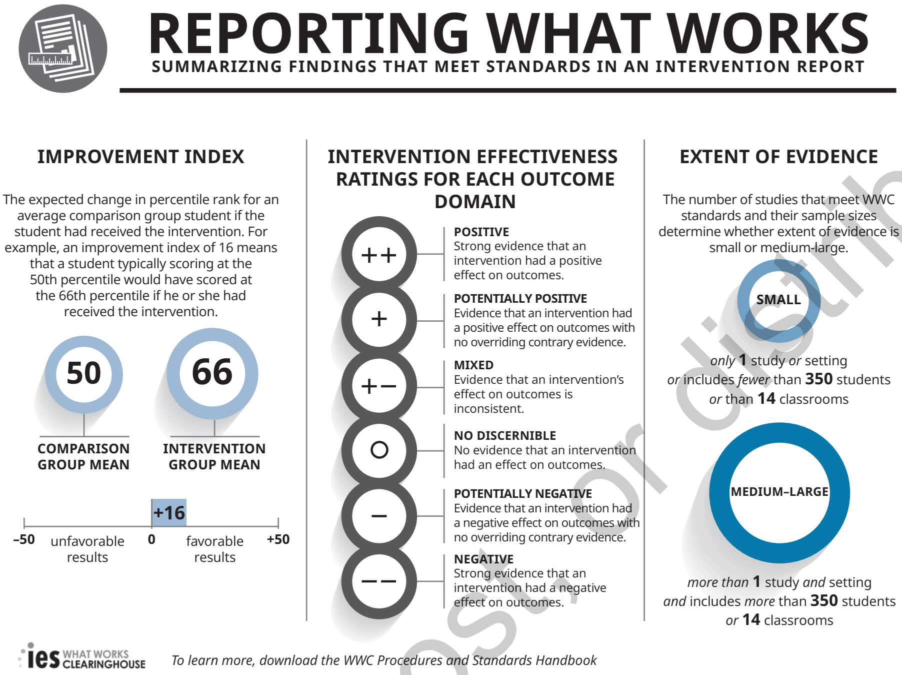
Source: What Works Clearinghouse. (n.d.c). *What is the WWC? A trusted source about what works in education*. Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/SWAT-What-is-the-WWC-v2_508.pdf

FIGURE 2.2 ■ What Works Clearinghouse Rating Process


HOW THE WWC RATES A STUDY
RATING GROUP DESIGNS
Are groups randomly assigned?
YES
NO
Is sample attrition high or low?
HIGH
LOW
Are groups similar before the intervention began?
YES
NO
Are there confounding factors or concerns with outcomes?
YES
NO
Are there confounding factors or concerns with outcomes?
YES
NO
MEETS WWC STANDARDS WITHOUT RESERVATIONS
MEETS WWC STANDARDS WITH RESERVATIONS
DOES NOT MEET WWC STANDARDS
Reviews of individual studies are the foundation of all WWC review products:
INTERVENTION REPORTS
SINGLE STUDY REVIEWS
QUICK REVIEWS
PRACTICE GUIDES
ies WHAT WORKS CLEARINGHOUSE
To learn more, download the WWC Procedures and Standards Handbook

Source: What Works Clearinghouse. (n.d.a). *How the WWC rates a study: Rating group designs*. Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_info_rates_061015.pdf

FIGURE 2.3 ■ What Works Clearinghouse Intervention Report Definitions



REPORTING WHAT WORKS

SUMMARIZING FINDINGS THAT MEET STANDARDS IN AN INTERVENTION REPORT

IMPROVEMENT INDEX

The expected change in percentile rank for an average comparison group student if the student had received the intervention. For example, an improvement index of 16 means that a student typically scoring at the 50th percentile would have scored at the 66th percentile if he or she had received the intervention.

INTERVENTION EFFECTIVENESS RATINGS FOR EACH OUTCOME DOMAIN

++ POSITIVE
Strong evidence that an intervention had a positive effect on outcomes.

+ POTENTIALLY POSITIVE
Evidence that an intervention had a positive effect on outcomes with no overriding contrary evidence.

+− MIXED
Evidence that an intervention's effect on outcomes is inconsistent.

○ NO DISCERNIBLE
No evidence that an intervention had an effect on outcomes.

− POTENTIALLY NEGATIVE
Evidence that an intervention had a negative effect on outcomes with no overriding contrary evidence.

−− NEGATIVE
Strong evidence that an intervention had a negative effect on outcomes.

EXTENT OF EVIDENCE

The number of studies that meet WWC standards and their sample sizes determine whether extent of evidence is small or medium-large.

SMALL

only **1** study *or* setting
or includes *fewer* than **350** students
or than **14** classrooms

MEDIUM-LARGE

more than **1** study *and* setting
and includes *more* than **350** students
or **14** classrooms

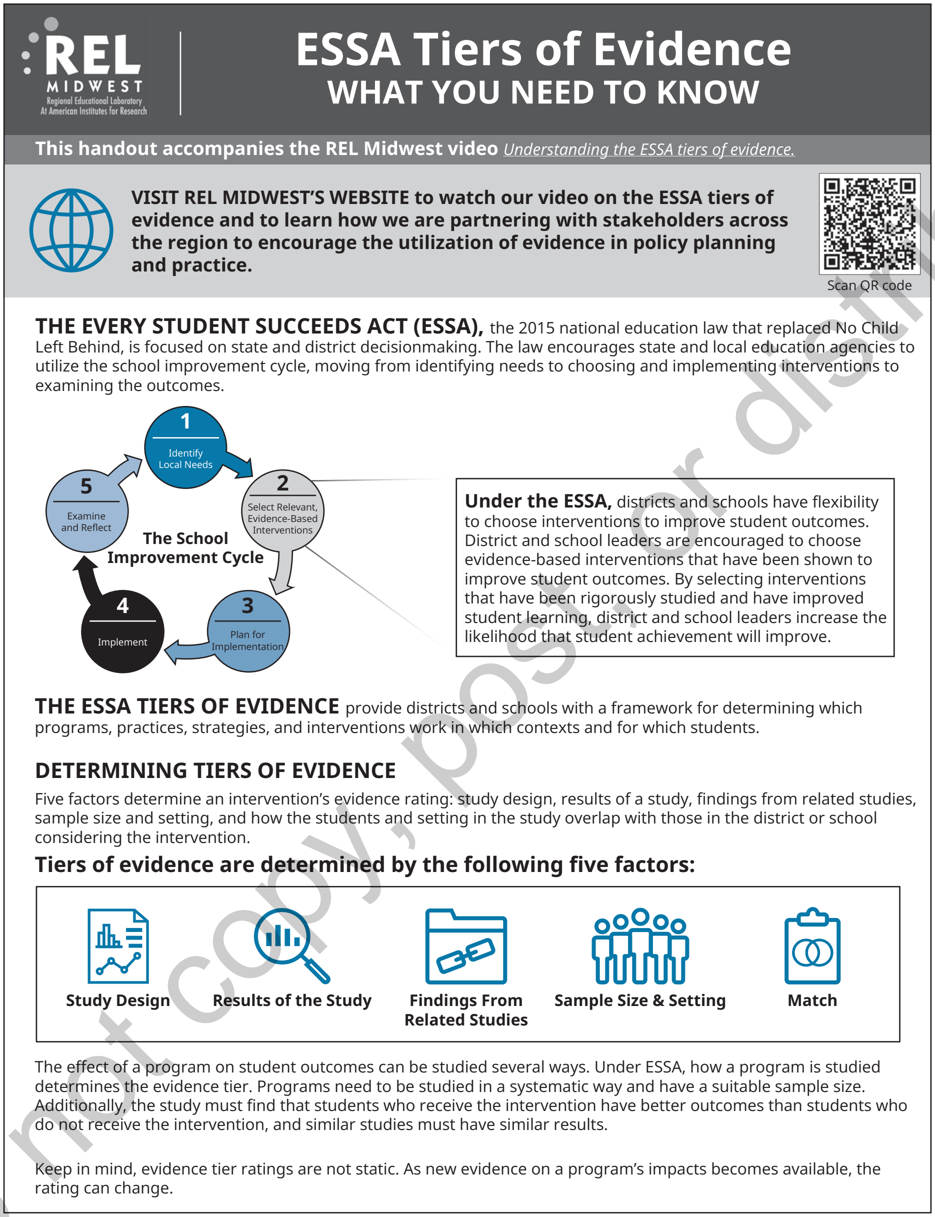
To learn more, download the WWC Procedures and Standards Handbook

Source: What Works Clearinghouse. (n.d.b). *Reporting what works: Summarizing findings that meet standards in an intervention report*. Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_info_reporting_061015.pdf

with the setting in which the program may be implemented. Note that the fifth factor is context-specific; that is, an organization considering an intervention or program should examine how closely their implementation setting, including the target audience, aligns with the setting(s) in which and population(s) with which the intervention was studied. Based on these factors, a study is assigned one of four tiers: strong evidence (Tier 1), moderate evidence (Tier 2), promising evidence (Tier 3), or demonstrates a rationale (Tier 4). See Figure 2.4 for additional information on how ESSA Tiers of Evidence are determined (REL Midwest, 2019).

The WWC intervention reports and their reviews of individual studies provide a resource for evaluators to examine what research has already been done, and its quality, on multiple topics and for practitioners to understand what evidence-based practices and programs exist in a given area.

FIGURE 2.4 ■ ESSA Tiers of Evidence



REL MIDWEST
Regional Educational Laboratory
At American Institutes for Research

ESSA Tiers of Evidence

WHAT YOU NEED TO KNOW

This handout accompanies the REL Midwest video *Understanding the ESSA tiers of evidence.*

VISIT REL MIDWEST'S WEBSITE to watch our video on the ESSA tiers of evidence and to learn how we are partnering with stakeholders across the region to encourage the utilization of evidence in policy planning and practice.

Scan QR code

THE EVERY STUDENT SUCCEEDS ACT (ESSA), the 2015 national education law that replaced No Child Left Behind, is focused on state and district decisionmaking. The law encourages state and local education agencies to utilize the school improvement cycle, moving from identifying needs to choosing and implementing interventions to examining the outcomes.

The School Improvement Cycle
1 Identify Local Needs
2 Select Relevant, Evidence-Based Interventions
3 Plan for Implementation
4 Implement
5 Examine and Reflect

Under the ESSA, districts and schools have flexibility to choose interventions to improve student outcomes. District and school leaders are encouraged to choose evidence-based interventions that have been shown to improve student outcomes. By selecting interventions that have been rigorously studied and have improved student learning, district and school leaders increase the likelihood that student achievement will improve.

THE ESSA TIERS OF EVIDENCE provide districts and schools with a framework for determining which programs, practices, strategies, and interventions work in which contexts and for which students.

DETERMINING TIERS OF EVIDENCE

Five factors determine an intervention's evidence rating: study design, results of a study, findings from related studies, sample size and setting, and how the students and setting in the study overlap with those in the district or school considering the intervention.

Tiers of evidence are determined by the following five factors:

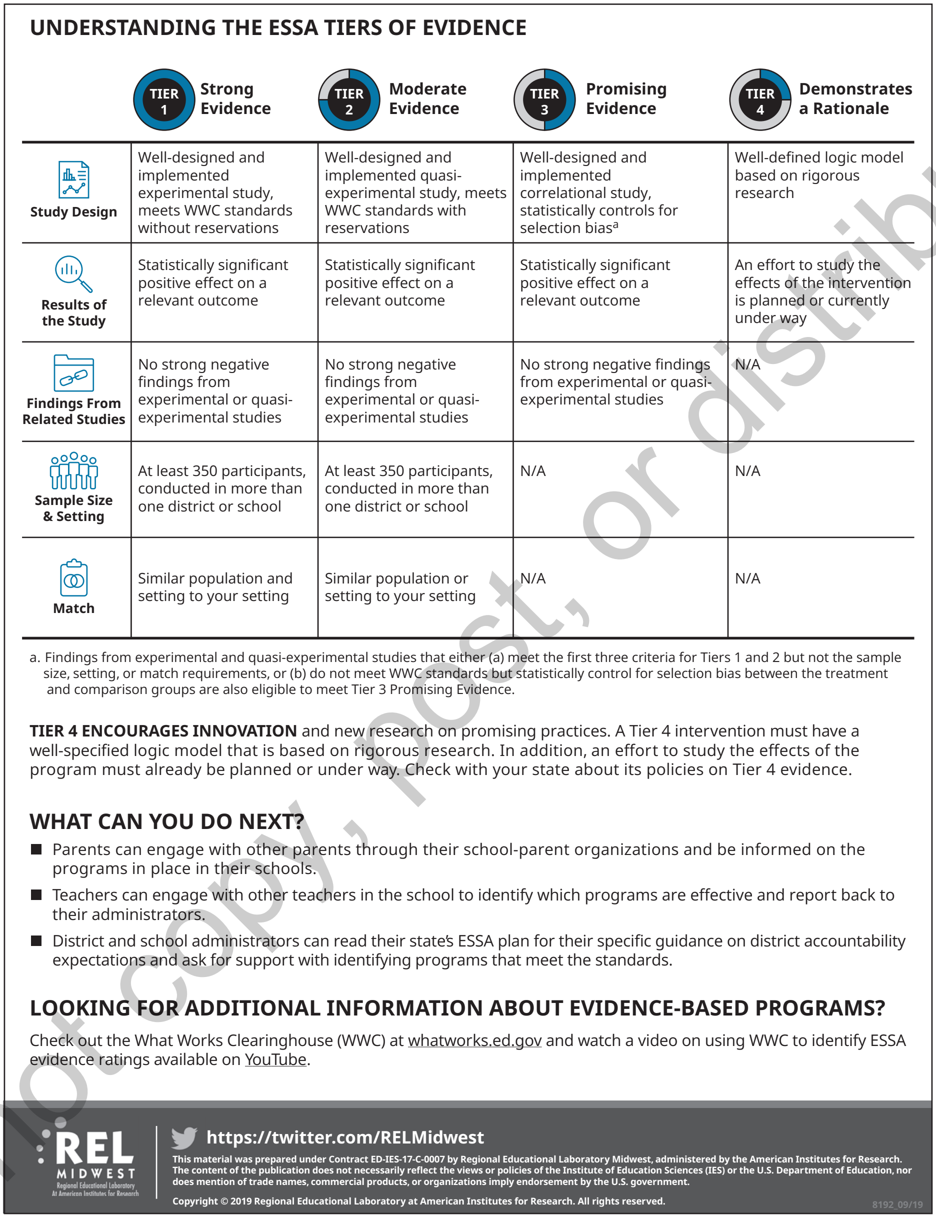
Study Design | **Results of the Study** | **Findings From Related Studies** | **Sample Size & Setting** | **Match**

The effect of a program on student outcomes can be studied several ways. Under ESSA, how a program is studied determines the evidence tier. Programs need to be studied in a systematic way and have a suitable sample size. Additionally, the study must find that students who receive the intervention have better outcomes than students who do not receive the intervention, and similar studies must have similar results.

Keep in mind, evidence tier ratings are not static. As new evidence on a program's impacts becomes available, the rating can change.

(Continued)

FIGURE 2.4 ■ ESSA Tiers of Evidence (*Continued*)



UNDERSTANDING THE ESSA TIERS OF EVIDENCE

	TIER 1 Strong Evidence	TIER 2 Moderate Evidence	TIER 3 Promising Evidence	TIER 4 Demonstrates a Rationale
Study Design	Well-designed and implemented experimental study, meets WWC standards without reservations	Well-designed and implemented quasi-experimental study, meets WWC standards with reservations	Well-designed and implemented correlational study, statistically controls for selection bias[a]	Well-defined logic model based on rigorous research
Results of the Study	Statistically significant positive effect on a relevant outcome	Statistically significant positive effect on a relevant outcome	Statistically significant positive effect on a relevant outcome	An effort to study the effects of the intervention is planned or currently under way
Findings From Related Studies	No strong negative findings from experimental or quasi-experimental studies	No strong negative findings from experimental or quasi-experimental studies	No strong negative findings from experimental or quasi-experimental studies	N/A
Sample Size & Setting	At least 350 participants, conducted in more than one district or school	At least 350 participants, conducted in more than one district or school	N/A	N/A
Match	Similar population and setting to your setting	Similar population or setting to your setting	N/A	N/A

a. Findings from experimental and quasi-experimental studies that either (a) meet the first three criteria for Tiers 1 and 2 but not the sample size, setting, or match requirements, or (b) do not meet WWC standards but statistically control for selection bias between the treatment and comparison groups are also eligible to meet Tier 3 Promising Evidence.

TIER 4 ENCOURAGES INNOVATION and new research on promising practices. A Tier 4 intervention must have a well-specified logic model that is based on rigorous research. In addition, an effort to study the effects of the program must already be planned or under way. Check with your state about its policies on Tier 4 evidence.

WHAT CAN YOU DO NEXT?

- Parents can engage with other parents through their school-parent organizations and be informed on the programs in place in their schools.
- Teachers can engage with other teachers in the school to identify which programs are effective and report back to their administrators.
- District and school administrators can read their state's ESSA plan for their specific guidance on district accountability expectations and ask for support with identifying programs that meet the standards.

LOOKING FOR ADDITIONAL INFORMATION ABOUT EVIDENCE-BASED PROGRAMS?

Check out the What Works Clearinghouse (WWC) at whatworks.ed.gov and watch a video on using WWC to identify ESSA evidence ratings available on YouTube.

REL MIDWEST Regional Educational Laboratory At American Institutes for Research

https://twitter.com/RELMidwest

This material was prepared under Contract ED-IES-17-C-0007 by Regional Educational Laboratory Midwest, administered by the American Institutes for Research. The content of the publication does not necessarily reflect the views or policies of the Institute of Education Sciences (IES) or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. government.

8192_09/19

Source: What Works Clearinghouse on IES website

QUICK CHECK

1. What experiments are considered the worst ethical violations in human history? How did they violate ethical principles?
2. What do the Tuskegee and Guatemala experiments have in common?
3. In comparing experiments such as the Holmesburg and MK-Ultra to experiments such as the Stanford Prison and Milgram, what are your thoughts on medical harm versus psychological harm?
4. How can AI be used to advance the field of evaluation? How can evaluation be used to foster the appropriate use of AI?
5. What does the What Works Clearinghouse tell us about interventions?

Finally, the AEA provides critical infrastructure and support for the field of evaluation and guiding principles for evaluators (see Chapter 4); core competencies for evaluation professionals; content- and methodology-focused topical interest groups as a means for evaluators to share ideas and collaborate; links to resources and evaluator blogs; professional development opportunities, including summer institutes and webinars, for evaluators to learn new skills; evaluator recognition; journals to disseminate best practices and professional advances; discussion forums; and events for evaluators to network, learn, and share, such as the annual meeting. AEA can be accessed at https://www.eval.org/

2.4 CHAPTER SUMMARY

In this chapter, the history of evaluation is discussed from two perspectives: development of the field of evaluation and development of research ethics that affect how evaluations are conducted. While evaluation as a human activity has been around as long as humans have walked on Earth, evaluation as a method of examining programs is rather new. Two primary events shaped the field of evaluation, namely the launching of Sputnik by the Russians in 1957 and the proliferation of social programs in the 1960s. Sputnik forced the United States to accelerate its space program and reexamine the ways in which scientists are prepared. The American education system, in particular, became a focus for improvement.

Investment in numerous social programs eventually prompted a focus on whether the programs were cost-efficient and cost-effective. Both created a need for evaluators of programs. Individuals from many fields came together to shape the field of evaluation, including psychologists, educators, and sociologists. During the 1970s and 1980s, universities began to offer courses in evaluation and the **American Evaluation Association** was created. AEA is an international professional association of evaluators focused on sharing approaches and methods.

Ethical guidelines in evaluation are based on research ethics. There is a disturbing national and international history of ethical violations in research. The experiments by Nazi Germany during World War II are perhaps the worst example of humans abusing humans in the name of research, though the United States government has also conducted numerous unethical experiments on humans, including the Tuskegee syphilis experiments and the sponsorship of widespread radiation experiments. There are also numerous examples of unethical treatment of human subjects by American researchers. Two of the best known are the Stanford prison experiments and the Milgram experiments.

Along with the history of evaluation and this history of ethical violations in research, common concerns of evaluators regarding program evaluation are presented, as well as infrastructure supports to address these concerns. Five concerns are explained: (1) the methods evaluators use to conduct evaluations, (2) the way in which results are communicated, (3) how these results are used for program improvement, (4) the extent to which findings can influence the policy process, and (5) how evaluation practice can be organized to address issues of design, reporting, use, and influence. Current issues in evaluation include the use of AI technologies and advancing the incorporation DEIA into evaluation practice. Finally, professional organizations, such as the AEA, and resources, such as the What Works Clearinghouse, are infrastructure supports that can aid evaluators in addressing, discussing, and building knowledge about some of these common concerns.

REFLECTION AND APPLICATION

1. In the chapter, the human radiation experiments conducted by the United States are introduced. The U.S. Department of Energy documented at least 425 such experiments, including a study conducted at Vanderbilt University in which over 800 pregnant women were given radioactive iron to test its absorption. A 1995 document by the U.S. Department of Energy summarizes these experiments (https://www.osti.gov/opennet/servlets/purl/16141769/16141769.pdf). Go to this document and choose one experiment; search the Internet to see if you can find additional information on this experiment.
   a. What was the purpose of the experiment?
   b. When was it conducted?
   c. Who participated in the experiment?
   d. Did the subjects know they were participants in a study?
   e. Was there any resolution, settlement, or apology as a result of the experiment?
2. Go to What Works Clearinghouse (https://ies.ed.gov/ncee/wwc/). Choose a topic and explore the research on this topic. How can evaluators use the WWC to address some of the concerns presented by Shadish in his "common threads" editorial?

KEY TERM

See Glossary for all definitions.

American Evaluation Association (AEA)