2

HISTORY OF EVALUATION

If you don't know history, then you don't know anything. You are a leaf that doesn't know it is part of a tree.

—Michael Crichton

Upon completion of this chapter, you should be able to

- Discuss the historical context of evaluation.
- Describe how the discipline of evaluation has evolved over the last 200 years.
- Identify important contributors to the development of the field of evaluation.
- Explain the history of research and evaluation with respect to ethics.
- 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. Eve made the decision to eat from the forbidden tree, evaluating the information that she had and obviously weighting one source more than another. Her information was conflicting and she did not foresee the consequences of her decision, but it was evaluative nonetheless. Some researchers look back further and place the roots of evaluation with evolutionary biology (Shadish, Cook, & Leviton, 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, 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 one thousand years later in 1115 BCE, the Chan dynasty began testing staff before they were hired; and over two thousand 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, 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 in order 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 in 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 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 (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 of using student test scores to evaluate educational programs remains in use today (Hogan, 2007). Due to the quantitative nature of evaluative systems through the mid-1800s, 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" on the next page for more information on the Cambridge-Somerville Youth Study (Cabot, 1940; McCord, 1978, 2002, 2003; McCord & McCord, 1959).

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,

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 four 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 ten years, with two- to three-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 (1940).

and was administered for the first time in 1969 (Vinovskis, 1998). NAEP has been used for 50 years and currently tests across 10 content areas in Grades 4, 8, and 12 (National Center for Education Statistics, 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).

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 in to 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 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 2019, 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 particular contributions of individuals to the field of evaluation, as well as a discussion of evaluation approaches, will be 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 merged. 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. The AEA now has about 7,300 members across more than 80 countries (see www.eval.org).

American Evaluation Association (AEA): an international professional association of evaluators focused on sharing knowledge of evaluation approaches and methods.

During the 1990s, the U.S. government increased funding for and amplified 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

TABLE 2.1 ⬢ Early Contributors to the Field of Evaluation

Evaluator	Employment (if applicable, as of 2019)	Degree Received From/Date	Degree/Field of Study
Donald T. Campbell	*Deceased 1996*	University of California, Berkeley (1947)	PhD, Psychology
Huey T. Chen	Mercer University	University of Massachusetts—Amherst[1]	PhD, Sociology
Thomas D. Cook	Northwestern University	Stanford University (1967)	PhD, Communications Research
Lee J. Cronbach	*Deceased 2001*	University of Chicago (1940)	PhD, Educational Psychology
Stewart I. Donaldson	Claremont Graduate University	Claremont Graduate University (1991)	PhD, Psychology
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
Ernest R. House	University of Colorado Boulder	University of Illinois at Urbana-Champaign (1968)	EdD, Education
Mark W. Lipsey	Vanderbilt University	Johns Hopkins University (1972)	PhD, Psychology
Mel M. Mark	Pennsylvania State University	Northwestern University (1979)	PhD, Psychology
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	Claremont Graduate University	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

(*Continued*)

TABLE 2.1 ⬢ (Continued)

Evaluator	Employment (if applicable, as of 2019)	Degree Received From/Date	Degree/Field of Study
Daniel L. Stufflebeam	*Deceased 2017*	Purdue University (1964)	PhD, Statistics and Measurement
William M. K. Trochim	Cornell University	Northwestern University (1980)	PhD, Methodology and Evaluation Research
Ralph W. Tyler	*Deceased 1994*	University of Chicago (1927)	Educational Psychology
Carol H. Weiss	*Deceased 2013*	Columbia University (1977)	PhD, Sociology
Joseph S. Wholey	University of Southern California	Harvard University[1]	PhD, Philosophy

1. Unable to locate year of degree.

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, will be reviewed in Chapter 4.

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 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–present)

Hogan describes the Age of Reform as the time when the first recorded evaluation took place. As mentioned above, higher education institutions in England and the United States began to use quantitative methods to evaluate students, 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, as 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 propelled the United States further into growth mode; however, conversations arose regarding accountability and the 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 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 new field. Professional organizations were formed, evaluation methodologies generated, methods of dissemination (such as journals) created, and university programs focused on producing evaluation professionals developed. Finally, the Age of Expansion and Integration continues to this day. Evaluation is now recognized as a profession and, further, infrastructure to support the field of evaluation has been developed. Yet evaluation is still a relatively young field and there are many opportunities for discovery and growth.

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 five-year period between 1939 and 1944.

Findings after 15 years as analyzed by McCord and McCord (1959) 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 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 upon 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 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 as to 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). Other researchers point to subsequent research on protective factors, social influences, and institutional influences (Welsh, Zane, & Rocque, 2017).

Sources: McCord (1978, 2002, 2003); McCord & McCord (1959); Vosburgh & Alexander (1980); Welsh et al. (2017).

QUICK CHECK

1. Who is considered the "Father of Evaluation" and why?
2. What event occurred in the 1950s that helped to 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, 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 done 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 will examine 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 on humans with deadly poisons (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 poor African American farmers 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. Over 1,300 people were intentionally infected with venereal diseases, including syphilis, 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 (Resnick, 2019). 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 in order 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, scan 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, Collins, & Brief, 1995).

Tearoom Trade. 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 encounters through field

notes while serving as the lookout, or "watchqueen," during these impersonal sexual encounters (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 in order 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. Humphreys's tactics have received much criticism, regardless of whether some believe his findings to be informative (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 two weeks, but was terminated after six days due to abusive conditions and psychological distress. While most agree that the experiment violated ethical standards, there is continuing discussion as to 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 of the unethical experiments 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 evaluators. The history of legislation related to human-subject protections and ethical conduct of researchers will be 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. 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 several months back 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 and 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 CURRENT ISSUES IN EVALUATION

There are many relevant and timely issues in evaluation that will be covered throughout the text. In this section, we will discuss the common issues in evaluation and infrastructure supports that address 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 to shape the field of evaluation, and these concerns still permeate every meeting of the American Evaluation Association. 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. In order 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.

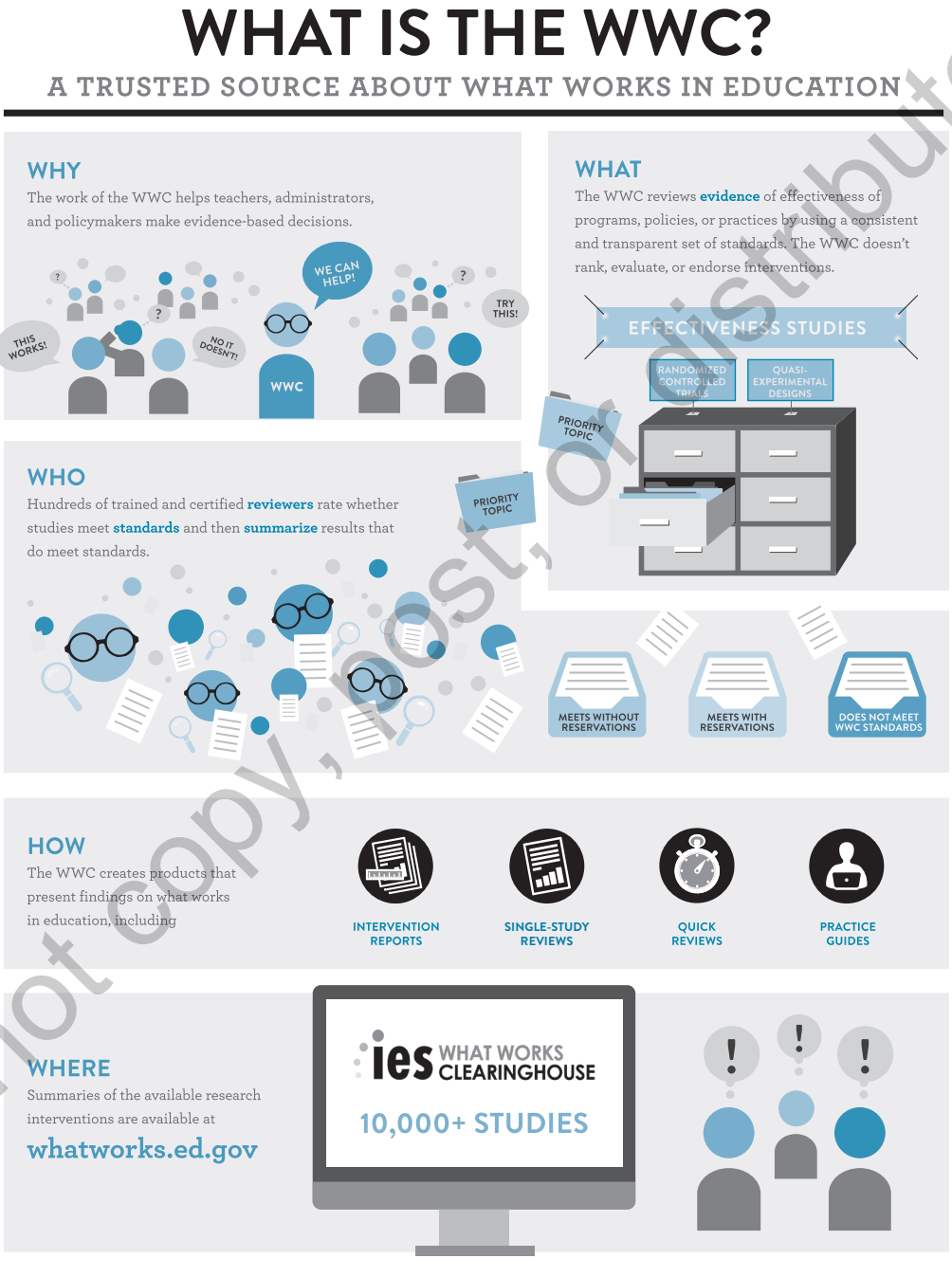
2.3.2 Resource Sharing and Dissemination

The Cochrane Collaboration is an international organization that provides synthesized research evidence around topics in health care. Cochrane 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/.

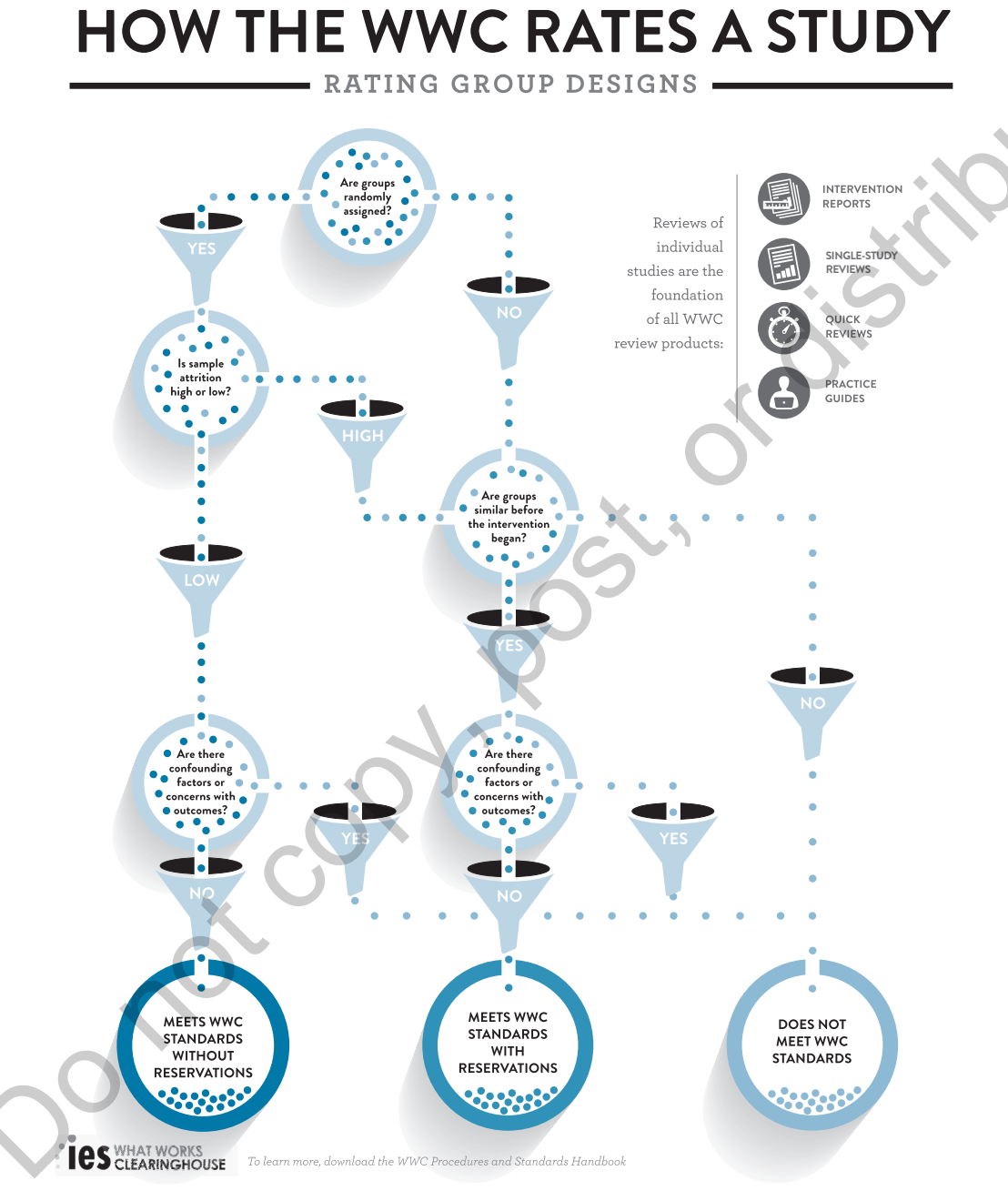
The What Works Clearinghouse (WWC) is a resource provided by the Institute of Education Sciences (IES). It was created in 2002 with the involvement of some of the same researchers who helped to start the Campbell Collaboration. The WWC includes evidence-based practices and programs in many education-related areas, including early-childhood education, literacy, behavior, and mathematics. The clearinghouse creates intervention reports through a rigorous review process based on rating studies according to a set of standards and then summarizes findings for studies that do meet standards. 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. The WWC provides a resource for evaluators to examine what research has already been done on topics and for practitioners to understand what evidence-based practices and programs exist on a given topic.

FIGURE 2.1 ⬢ What Works Clearinghouse



Source: What Works Clearinghouse on IES website at https://ies.ed.gov/blogs/ncee/post/five-reasons-to-visit-the-what-works-clearinghouse

FIGURE 2.2 ⬢ What Works Clearinghouse Rating Process



Sources: What Works Clearinghouse on IES website at https://ies.ed.gov/blogs/ncee/post/five-reasons-to-visit-the-what-works-clearinghouse

QUICK CHECK

1. What experiments are considered the worst ethical violations in human history? Who conducted the experiments? How many people were affected?
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. Explain the five concerns of evaluators that Shadish summarized from the evaluation literature.

Finally, the American Evaluation Association provides critical infrastructure and support for the field of evaluation. AEA provides 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 was 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 the 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. In 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 world 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. 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 were 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 the 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?