2

Assessing the Quality of Evidence

When is evidence sufficient to support a claim about how efficiently, effectively, or equitably a program is implemented or "working?" What constitutes high-quality data and rigorous evaluation research? Whose criteria for judging the rigor of research should be applied? There are many questions that arise once we take a closer look at the factors surrounding the production of the evidence that could be used to inform decision-making in government.

In this chapter, we address the basic issue of how to assess the quality of evidence, and offer widely accepted criteria for judging the quality of evidence, as well as evaluation and research study findings. We describe the differences existing across producers and the variety of users, e.g., managers and beneficiaries, regarding what constitutes sufficient rigorous evidence, and offer guidance on educating relevant stakeholders on how to assess the quality of evidence.

When Is Evidence Compelling to Inform Public Policy?

Government auditors have been collecting data about government operations for many decades, and have developed a rule of evidence that they have carefully crafted and refined over time to ensure that the evidence they provide to support each claim they make is credible and compelling. Their statement about the quality of evidence appears in the *Government Auditing Standards* (2018) and it offers an extremely useful and defensible tool for anyone operating in the public policy arena. They state that "Auditors must obtain **sufficient, appropriate** evidence to provide a reasonable basis for addressing the audit objectives and supporting their findings and conclusions" (GAO 2018, p. 179).

Over time, the way federal government auditors have explained the terms "appropriate" and "sufficient" have changed, but the recent language in the *Quality Standards for Inspection and Evaluation (2012)* based on the GAO rule and published within the Federal Inspectors General community summarized the goal succinctly, as: "Evidence supporting inspection findings, conclusions, and recommendations should be **sufficient, competent, and relevant** and should lead a reasonable person to sustain the findings, conclusions, and recommendations" (PCIE Bluebook 2012, p. 12).

Competent

The criteria of competent, relevant, and sufficient work quite well with any evidence claim that might be made with data collected or generated in the public arena. Competence is perhaps the most understandable criterion for researchers and evaluators to address, for it refers to the use of generally accepted research methods, and might be worded as: Was the appropriate methodology used competently by well-trained professionals to collect the data and generate the evidence?

There are fairly clear standards and expectations shared across the natural and social sciences on what constitutes competent application of research methods. The general criteria of validity and reliability are most often used, although the expectations for applying these criteria were originally developed to apply to quantitative data. As the availability and use of qualitative data, i.e., words rather than numbers, has increased, the need for more appropriate criteria to assess the quality of qualitative data has increased as well. Table 2.1 offers generally accepted definitions of the multiple dimensions of validity that researchers take into account when collecting and analyzing both quantitative and qualitative data.

TABLE 2.1 ● Criteria for Assessing the Quality of Quantitative and Qualitative Data and Research and Evaluation Findings

Criteria	For Quantitative Data	For Qualitative Data
Measurement Accuracy	Measurement Validity: *To what extent are researchers/ evaluators accurately measuring what they really intend to measure?*	Authenticity: *To what extent are researchers/evaluators capturing the voice and meaning of the observed concepts in the way they intended?*
Measurement Processes	Measurement Reliability: *To what extent will the way in which measurement occurs be expected to produce similar results on repeated observations*	Auditability: *How clear and transparent are the procedures researcher/evaluators use to generate data? To what extent do they clearly explain how, when,*

Criteria	For Quantitative Data	For Qualitative Data
	of the same condition or event? Are data collected and entered consistently? Would others obtain the same answer if they repeated the question or data collection task?	*and in what contexts they generated data via asking questions and/or making observations?*
Causal Claims	Internal Validity: *To what extent are researchers/evaluators able to establish that there is a causal relationship between a specified cause and effect?*	Confirmability of Inferences: *To what extent do the data provided support claims made about explanations for the occurrence of observed effects? (Note that qualitative research methods, such as process tracing and Qualitative Comparative Analysis (QCA), are sometimes employed to search for the causes of observed outcomes, and may provide explanations that should be confirmable via the evidence provided.)*
Generalizability	External Validity: *To what extent are researchers/evaluators able to generalize from the results* to *groups or contexts beyond those being studied?*	Transferability and/or Fittingness: To what extent are *findings, such as about processes, deemed relevant to be applicable in other locations and other times? (Note the context should be similar to that in which the research was undertaken).*
Statistical Inferences	Statistical Conclusion Validity: *To what extent do the numbers researchers/evaluators generate in a sample accurately measure the magnitude of a factor or an effect, or strength of a relationship in the population from which the sample was drawn?*	Not Applicable
Multicultural Validity[a]	**(Note: Multicultural Validity Applies to the Entire Evaluation Process For Both Quantitative and Qualitative Data Collection and Analysis)** Multicultural Validity: *To what* ***extent*** *have researchers/evaluators respected and taken into account the following elements during design, data collection, and analysis: history, location, power, relationship, voice, time, return, plasticity, and reflexivity?*	

(Continued)

TABLE 2.1 ● *(Continued)*

	History of place, people, program, and evaluation's role, including knowledge of cultural heritages and traditions, and their evolution over time. *Location* includes cultural contexts and affiliations of evaluators and subjects, including theories, values, meaning-making, and worldviews; and recognizing the multiple cultural intersections at individual, organizational, and systems levels and the geographic anchors of culture in place. *Power* entails understanding how privilege is attached to some cultural signifiers and prejudice to others; and addressing equity and social justice, and not perpetuating condescension, discrimination, or disparity. *Relationship* entails maintaining strong connections among the evaluation, program or policy, and the community, and establishing trust and maintaining accountability to the community with respect and responsibility. *Voice* entails clarifying whose perspectives are magnified and whose are silenced, and mapping inclusion and exclusion or marginalization. *Time* involves attending to the rhythm, pace, and scheduling and to the participants' vision of past and future and involves considering longer impacts and implications—positive or negative. *Return* entails focusing attention both during and after the evaluation process regarding how the evaluation and/or the persons who conduct it return benefit to those studied and the surrounding community and ensuring the evaluation is not exploitive. *Plasticity* entails allowing the evaluators and the evaluation design, processes, and products to respond and adapt when receiving new information, and change in response to new experiences since culture is fluid, not static. *Reflexivity* entails evaluators being reflective of how their own values and world views affect their practice, as well as their evaluation design, processes, and products.

[a] Explanation of multicultural validity is adapted from Kirkhart (2013, p. 150).

The ways that quantitative and qualitative data and methods may be assessed are aligned in Table 2.1 to facilitate presentation, but in fact, the intent and rationale underlying the different criteria to apply are actually quite different. There are a range of values and preferences on how to learn about the physical and social world held by researchers. Differences in interpreting and applying the criteria between researchers employing qualitative research methods, such as in-depth interviewing and observation, and researchers using quantitative data gathered via surveys and administrative data collection in agencies may be great. For example, many qualitative researchers reject the expectation that they would hear the same answer if they interviewed a participant a second time, thus the notion of auditability does not convey the same expectation as reliability—which does entail the expectation of replicability (Mason 2018).

Accuracy of measurement is a shared goal of both quantitative and qualitative researchers, although it may be approached slightly differently. Researchers start with a concept, such as maternal health, and identify measurement procedures that they can use to operationalize the more abstract concept into empirically observable indicators. Quantitative data such as blood pressure counts might be used to assess in part pregnant women's well-being, and interviewers might ask the expectant mothers to self-report how they feel and what daily activities they find it difficult to complete. In some arenas there are quantitative data commonly used to operationalize targeted measures, for example, standardized reading exams are used to measure 4th grade student reading abilities. For other targeted measures, there may not be commonly accepted measures so researchers develop new measures and are expected to support their choices.

Evaluators and researchers frequently draw upon long-standing strategies used in psychology to demonstrate the accuracy and adequacy of measures, a process called validation. In order to assess and convey the accuracy of the measures they employ, researchers frequently rely on the views of the relevant subject matter experts who can attest to the validity of the empirical indicators used (called content validation). In addition, they may rely on criterion validation, where they test the empirical relationship between a new measure and commonly accepted indicators of the attribute of interest. For example, the Body Mass Index (BMI) is the most often used measure of obesity and a new measure may be tested to see if it correlates strongly with the BMI to assess that it may also provide a good measure of obesity. Predictive validation entails testing the extent to which the measure forecasts future performance on the attribute of interest. For example, the validity of GRE scores may be tested by measuring how well they predict the grades of doctoral students once they are in a doctoral program, and LSAT scores may be tested as to how well they predict performance in law school. Consequential validation also entails measuring the association of the measure with intended (and perhaps unintended) outcomes of using the measure. For example, one might use the percentage of babies immunized (output) to predict the rate of polio in a region (outcome) to demonstrate the accuracy of the immunization rate as a measure.

All evaluators and researchers are expected to describe the measurement processes they employ. For quantitative data collection, measurement reliability entails demonstrating that both reliable measures and reliable measurement processes are used. A measure is deemed reliable if the operation employed consistently measures the same phenomenon, for example, is a question asked in the same way? And reliable measurement entails consistently recording data with the same decision criteria, across time and location.

Most qualitative researchers do not claim that a measurement process will yield exactly the same answers if replicated. However, they are expected to demonstrate the rigor of the processes they employ to generate data that are typically elicited from participants by recording exactly how, when, and where they ask questions and/or record observations. They are also expected to clarify and transparently describe all steps taken to identify themes from qualitative data they analyze. The criterion of auditability entails making transparent and clear the evidence trail, in other words, documentation that shows how data were generated or collected, and analyzed.

Causation is approached quite differently by researchers who bring different world views, and different disciplinary backgrounds to their work (Shadish, Cook, and Campbell 2002). Many quantitative researchers, including most economists, believe that if a study is designed well with adequate controls, it is possible to determine whether an intervention or treatment (e.g., a policy, program, external event, regulation, management action) caused an intended outcome (e.g., a gain in reading scores, a drop in unemployment, a change in infant mortality) and in what magnitude. Typically, social scientists assert that to conclude that the "cause" had the desired "effect," they must ascertain that (1) the cause preceded the effect in time; (2) the change in the cause can be linked to the change in the effect; and (3) no plausible other factors could have caused the change we observe in the effect. For example, if these conditions are met, higher levels of math achievement found in children taught with a new curriculum are then said to be caused by the curriculum. The rub is that researchers need to ensure that there are no other plausible factors responsible for the math gain, and that there are no methodological weaknesses in their approach that could have led them to conclude a causal impact when there was not one. Possible alternative explanations for measured effects are frequently numerous. It is critical that plausible "causal factors" that were not amenable to measurement in an evaluation are at least identified when discussing findings.

An evaluation or research study is typically designed to describe the magnitude of a causal impact by minimizing differences between the treatment group and the control group (which does not receive the treatment) (Peck 2020). The preferred choice for most researchers is to randomly assign the treatment and the control group from the same population, i.e., an experimental evaluation also referred to as a Random Control Trial (RCT). If random assignment is not possible due to ethical or logistical obstacles, such as when implementing a new clean air regulation, a research design that constitutes a comparison group that is extremely similar to the treatment group through advanced econometric techniques such as Propensity Score Matching or a Regression Discontinuity Design may be used to measure causal effects.

Measuring the average impact of a treatment through an experimental evaluation such as an RCT provides a causal description, but will typically not offer sufficient contextual information to explain why and how the observed impact was produced by the treatment (Shadish, Cook, and Campbell 2002, p. 12). With more simple interventions, such as teaching participants a skill, it may be that causal description is sufficient. However, with interventions that contain multiple components, such as a curriculum with multiple pieces for teachers to use, or a new case management approach that customizes services to support alienated youth, one rigorous design that measures changes in treatment and control (or comparison) groups may not provide the explanation needed as to which elements of the services worked for whom.

In order to answer how and why questions regarding the causes of effects, qualitative research methods are typically employed to help flesh out causal explanations. For example, qualitative data collections methods such as interviews might be used to accompany RCTs, and/or approaches such as process tracing and Qualitative Comparative Analysis may be used to describe the causes of observed outcomes, and may provide explanations that should be confirmable via the evidence provided. With explanatory claims supported with qualitative data, the conceptualization of unraveling causation is different from the notion of causal inference based on quantitative data, again reflecting the differing mental models and assumptions held by the researchers.

The objectives for generalizing results of studies also differ across quantitative and qualitative evaluation and research. For research based on quantitative data, external validity refers to the ability to apply a study's results to groups or contexts beyond those being studied. For example, if evaluators studied safety violations at a sample of 5 nuclear power plants in the United States, to what degree could the study's results be generalized to all US nuclear power plants? And how confidently could evaluators generalize survey results about purchases of healthy foods in high school cafeterias to the universe of high school students in the United States from a stratified random sample of 500 high schools across the nation? Typically, external validity is wedded to sampling, so that samples are carefully selected to be representative of the population from which they are drawn, and then statistical inferences are made from the sample values to other units in the population.

With research based on qualitative data, research objectives are typically to provide more nuanced explanations than obtainable with quantitative methods, and not to construct inferences that could be extrapolated to a large number of other locations. Rather than aspiring to generalize specific findings with some level of confidence, claims about processes, such as effective mentoring techniques, are offered as potentially transferable to other settings—as long as the other contexts are sufficiently similar to that in which the research was undertaken. The claims are offered as transferable only if there is a fit with the other contexts. Contextual factors that support the process or mechanism that appears to be effective are needed in the new context though, such as enthusiastic teachers or counselors who are interested in adopting a new mechanism or practice.

As noted above, statistical inferences are simply not offered in qualitative research. With quantitative research, statistical conclusion validity refers to how well the research as designed and implemented permits generalizing estimated values from the sample to the population from which the sample was drawn. Typically quantitative studies are designed to provide a sufficiently large and representative sample to detect observed outcomes. For example, a proposed design and analysis approach should be capable of detecting differences in reading achievement over three months between children taught with a new curriculum approach versus children taught with the traditional approach. There is always the possibility that methodological weaknesses in application of a statistical technique may have reduced (or increased) the likelihood of finding compelling differences between comparison groups, or evidence of important predictors of desired outcomes. For example, there could have been a great amount of attrition from a treatment group, thus those left for measurement of outcomes might be more highly motivated, presenting a critical rival explanation for their improved performance.

Interpreting the results of statistical tests is also a practical challenge, given the historic use of preset thresholds of statistical significance to make stronger claims than the tests are capable of demonstrating. Historically statistical hypothesis testing has been widely used to test the "statistical significance" of findings in quantitative research using preset and clear probability levels, e.g., *p*-values of .05 and .01, to make yes/no decisions on whether the hypothesis of no effect or no difference was to be rejected. In 2016, the American Statistical Association (ASA) provided guidance about the use of statistical significance and set the expectation that traditionally used *p*-values such as .05 were not appropriate to make yes/no decisions on hypotheses. The ASA's Statement specifies that *p*-values can indicate how incompatible the data are with a specified statistical model, but should not be used to measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. The ASA specified that scientific conclusions and business or policy decisions should not be based solely on whether a sample *p*-value passed a specific, preset threshold (see *The American Statistician*, Volume 70, 2016—Issue 2 Statement on P-Values).

The ASA policy on *p*-Values has been accepted and implemented very slowly across social science disciplines, with social psychologists presenting early adopters, but in other arenas, including economics (as of this writing) practice has not substantially changed (see *Basic and Applied Social Psychology*, Volume 37, 2015, Issue 1). The implications of the new normal in statistical significance testing include: preset *p*-values should not be used to test a finding's "significance" or correctness at all, nor to test the significance of means with a yes or no decision on their separation solely using the *t* test; the specific probability value of a statistic such as *t* should be provided and discussed, e.g., .062 or .049; and the probability values should be discussed and combined with other information, e.g., effect sizes, to support claims and conclusions (Wasserstein 2016). The use of confidence intervals around point or difference estimates is still promoted, however.

And lastly, the criterion that all data and findings have multicultural validity is extremely important. The principle that evaluators should be culturally competent has been discussed for several decades, and presents an aspirational value for evaluators. As the American Evaluation Association Statement on Cultural Competence states: "Cultural competence is not a state at which one arrives; rather, it is a process of learning, unlearning, and relearning. It is a sensibility cultivated throughout a lifetime. Cultural competence requires awareness of self, reflection on one's own cultural position, awareness of others' positions, and the ability to interact genuinely and respectfully with others. Culturally competent evaluators refrain from assuming they fully understand the perspectives of stakeholders whose backgrounds differ from their own" (see the Statement at https://www.eval.org/ccstatement).

Evaluation thought leaders in the United States have developed culturally responsive evaluation (CRE) concepts and frameworks intentionally to address racial and ethnic inequities within evaluation practice in the United States (see Hood, Hopson, and Kirkhart 2015). As Rodney Hopson defines it: "CRE is a theoretical, conceptual and inherently political position that includes the centrality of and [attunement] to culture in the theory and practice of evaluation. That is, CRE recognizes that demographic, sociopolitical, and contextual dimensions, locations, perspectives, and characteristics of culture matter fundamentally in evaluation" (Hopson 2009, p. 431).

In addition to those advocating use of CRE frameworks, a variety of evaluation thought leaders, including Jennifer Greene, David Fetterman, Ernest House, and Donna Mertens, advocate for evaluators to be sensitive to and responsive to cultural differences and inequities. There is a range of views among evaluation theorists and practitioners on when and how to use findings from evaluation work to advocate for policy change. Data collected in a culturally sensitive manner are needed to call out power imbalances and inequities, and such structural issues may not even be visible if evaluators do not take such measures. The bottom line is that regardless of how findings might be used, the evidence should be responsive and reflective of culture.

Multicultural validity is like a prism that should be used to view all decisions regarding the evaluation or research process, thus it affects the quality of evidence that is generated via any approach, and the perceived competence of quantitative and qualitative data and findings. As noted above, being culturally competent is not a static and easily achievable goal, but should be an aspirational goal of evaluators in each new context. Evaluators should assess and report on what they learned about each context, and what actions they took to ensure the process and findings were appropriate and reflective of the participants—whether evaluating a job training program for underemployed youth in Chicago or an economic empowerment program for women in Honduras. Similar to goals of human-centered design, it is critical to actively engage intended beneficiaries of any policy or program both in design and in evaluation to ensure the voice and lived experience of those in the community are taken into account. The clearest path to ensuring that any evaluation findings are duly reflective of a cultural context is to involve those community members most affected by a policy or program in the planning, design, and implementation of an evaluation. And importantly, the data and findings should be vetted so that they are viewed as appropriate and fitting by those affected by policy decisions the evidence is used to inform.

Beginning with Donald Campbell in the 1960s, social scientists have framed many limitations that may affect the validity and reliability of research and evaluation findings. Appendix 2.1 presents an inventory of typical limitations that should be acknowledged and addressed when appropriate in evaluation work. While many of the issues or threats to the credibility of our described claims were originally raised to question quantitative findings, most are appropriate to ask about qualitative data and findings as well.

Relevant and Sufficient

As enumerated above, there are many criteria to address to ensure that evidence from evaluation work is deemed competent, and the design and collection strategies employed are viewed as rigorous. Conceptualization of the different dimensions of validity is fairly similar across the social sciences, with some differences regarding how causation may or may not be established—an issue especially pertinent when trying to measure policy and program impact. Professional standards can help but will not dictate the "correct" methodological choices in each situation. Rigor is always affected by the resources available to collect and analyze data needed to answer the evaluation questions addressed. And producing "compelling" data and findings is more difficult when evaluating interventions in the field without much control over the context. Transparency and humility when conveying information about design decisions, measurement, data, and inferences are key to enhance perceptions of the rigor of evidence generated.

While the goal for evaluators and researchers is to produce convincing and understandable evidence to inform decisions and policymaking, the relevance and sufficiency of their work is judged by their audiences. Relevance refers to the extent to which the data and findings have a logical and clear-cut relationship with the issue or question being addressed. For example, if the impact of a job training program is being assessed, typically measures of the employment and earnings outcomes for participants are pertinent. And if the quality of services provided by a Department of Veterans Affairs medical center is being evaluated, data on medical outcomes are pertinent. Judgments will be made by audiences as to how relevant findings are to answer the questions that were raised.

One critical decision made by evaluators to ensure that the evidence they provide actually answers the key evaluation questions is that they employ the most appropriate research methods. Table 2.2 arrays the diverse sort of questions that might be framed and the more appropriate evaluation

TABLE 2.2 ● Match Designs and Data Collection Methods to the Evaluation Questions

Evaluation Objective	Illustrative Questions	Evaluation Design
#1: Describe program activities	• Who does the program affect—both targeted organizations and affected populations? • What activities are needed to implement the program (or policy)? By whom? • How extensive and costly are the program components? • How do implementation efforts vary across delivery sites, subgroups of beneficiaries, and/or across geographical regions? • Has the program (policy) been implemented sufficiently to be evaluated?	• Performance measurement • Exploratory evaluations • Evaluability assessments • Multiple case studies
#2: Probe implementation and targeting	• To what extent has the program been implemented? • When evidence-based interventions are implemented, how closely are the protocols implemented with fidelity to the original design?	• Multiple case studies • Implementation or process evaluations • Performance audits • Compliance audits

(Continued)

TABLE 2.2 ● *(Continued)*

Evaluation Objective	Illustrative Questions	Evaluation Design
	• What key contextual factors are likely to affect the ability of the program implementers to produce the intended outcomes? • What feasibility or management challenges hinder successful implementation of the program? • To what extent have activities undertaken affected the populations or organizations targeted by the regulation? • To what extent are implementation efforts in compliance with the law and other pertinent regulations? • To what extent does current program (or policy) targeting leave significant needs (problems) not addressed?	
#3: Measure program impact	• Has implementation of the program produced results consistent with its design (espoused purpose)? • How have measured effects varied across implementation approaches, organizations, and/or jurisdictions? • For which targeted populations has the program (or policy) consistently failed to show the intended impact?	• Experimental designs, i.e., random control trials • Difference-in-difference designs • Propensity score matching • Statistical adjustments with regression estimates of effects • Multiple time series designs • Regression discontinuity designs

Evaluation Objective	Illustrative Questions	Evaluation Design
	• Is the implementation strategy more (or less) effective in relation to its costs? • Is the implementation strategy more cost-effective than other implementation strategies also addressing the same problem? • What are the average effects across different implementations of the program (or policy)?	• Cost-effectiveness studies • Benefit–cost analysis • Systematic reviews • Metaanalyses
#4: Explain how and why programs produce intended and unintended effects	• How and why did the program have the intended effects? • Under what circumstances did the program produce the desired effects? • To what extent have program activities had important unanticipated negative spillover effects? • What are unanticipated positive effects of the program that emerge over time, given the complex web of interactions between the program and other programs, and who benefits? • For whom (which targeted organizations and/or populations) is the program more likely to produce the desired effects? • What is the likely impact trajectory of the program (over time)?	• Impact pathways and process tracing • Contribution analysis • Qualitative Comparative Analysis (QCA) • Nonlinear modeling, system dynamics • Configurational analysis, e.g., qualitative case analysis • Realist-based synthesis

(Continued)

TABLE 2.2 ● (Continued)

Evaluation Objective	Illustrative Questions	Evaluation Design
	• How likely is it that the program will have similar effects in other contexts (beyond the context studied)? • How likely is it that the program will have similar effects in the future?	

designs that could be employed to address them (see Newcomer, Hatry, and Wholey 2015). There are choices, and different designs require differing amounts of resources and skills, but the key is that appropriate comparisons and data are provided to answer the questions. Evaluators should clearly state why the design and data collection methods they used were appropriate. Appendix 2.2 provides a checklist to apply to evaluation work to assess the extent to which the evaluators provide enough information in order for audiences to judge the competence and relevance of their findings.

Decision-makers are the judges regarding how sufficient data and findings are to answer the questions they have. They judge whether or not there is enough evidence to support the findings and conclusions related to the questions that they want answered. What constitutes enough evidence? Judgments about how much evidence is enough are affected by the expectations, professional training, and values of the audiences for evaluation and research work. When making impactful decisions that affect great numbers of people and/or budgetary allocations, it is likely that more evidence is required, but how much more is again a judgment call.

Different Audiences, Different Judgments About the Quality of Evidence

Audiences for evidence about government performance and results bring to bear different professional standards and norms regarding what constitutes competent, relevant, and sufficient evidence. Evaluating the implementation and impact of complex public policies and programs entails making judgments by the evaluators, and by diverse audiences as they weigh evaluation findings. For example, lawyers, accountants, engineers, data scientists, economists, and anthropologists bring divergent professional norms, world views, and expectations to their assessment of evidence.

Even within disciplines and arenas of professional practice, there may be disagreements about preferred approaches and research methods. For example, for impact evaluation—especially when evaluating the impact of international development programming—there is controversy regarding the role of RCTs, i.e., randomized experiments, versus other sorts of evaluation designs in generating rigorous evidence. Some supporters argue that RCTs are the gold standard and present the only design that is capable of yielding sufficiently rigorous evidence, while others disagree (for example, see Pawson 2013).

Many government agencies and foundations across the world have been involved in reviewing social science research and evaluation studies to promote the use of stronger evidence to inform decision-making in the public arena, and they typically assess the rigor of the evidence they provide. The sponsors provide websites and online portals which house evaluations undertaken that describe specific interventions, their impact on specific outcomes, and their target populations and locations. The stated objectives of these clearinghouses include: providing a searchable database of programs and practices organized in a way that scholars and practitioners can search for programs of relevance to them; highlighting the most effective interventions to bring their evidence to larger audiences, especially practitioners, such as the What Works Website sponsored by the US Department of Education (see https://ies.ed.gov/ncee/wwc/); reviewing and synthesizing existing research on a topic to provide recommendations on what works and to what extent through systematic reviews on websites hosted by the Cochrane and the Campbell Collaborations (see https://www.cochrane.org/evidence and https://campbellcollaboration.org/); and, sometimes, calculating the overall effect of a program distilled in a single effect size using quantitative metaanalytical approaches. Some of the clearinghouses even provide guidance on how to implement the promising interventions they list, such as the California Evidence-Based Clearinghouse for Child Welfare (see https://www.cebc4cw.org/).

To address the quality of the evidence provided, the clearinghouses each use their own criteria and weighting schemes to rank the programs or studies they review. The criteria the various clearinghouses employ include many of the dimensions of validity discussed above for quantitative data, although they rarely address multicultural validity. It is rare for any clearinghouses to post studies based on qualitative data. Some of the clearinghouses only include studies that used RCTs, for example, and others assign studies into ordinal categories based on the designs used. For example, the Clearinghouse for Labor Evaluation and Research categorizes studies into three groups based on the design used, and only RCTs and interrupted time series designs garner the top category (High) versus the other two categories of Moderate and Low (see https://clear.dol.gov/). Application of criteria sometimes result in different ratings across clearinghouses, in fact, one early study found that of a random sample of 100 programs rated by more than one clearinghouse, 42% were inconsistently rated by the multiple sites to some degree (Means et al. 2015), though consistency improved in recent years.

The unit of analysis for rating evidence varies across the clearinghouses. Some clearinghouses rate individual studies, while others rate programs, or interventions. For example, CrimeSolutions.gov (under the National Institute of Justice) categorizes individual studies into classes ranging from 1 for low quality to 5 for high quality and then creates a cumulative measure from all of the studies pertaining to a specific program to produce a final evidence rating for each intervention. Similarly, the Cochrane Collaboration's primary focus is on metaanalysis, and the average outcome for an intervention is provided, with less emphasis on individual studies.

While most of the publicly funded clearinghouses have been established by federal agencies, a few have been state-funded. The state of Washington was a pioneer in establishing a public policy institute in 1983 to conduct research "to improve the ability of the Washington state legislature and other Washington state policymakers to make sound, evidence-based policy decisions" (see https://medium.com/data-labs/washington-state-institute-for-public-policy-wsipp-c91d7e40b8fd). The Washington State Institute for Public Policy is especially known for their work in carrying out cost–benefit analyses of publicly funded programs. For example, they provide cost–benefit analyses that calculate the benefit–cost ratios for a number of nonprofit grantees that are serving the same subpopulation to achieve the same specified outcome to show the relative cost-effectiveness of the providers. Given the complexity of the social problems addressed and questions about the comparability of overhead costs across nonprofits, the vulnerability of the rigor of benefit–cost ratios as the key piece of evidence to compare across providers may be open to question.

Promoting Shared Understanding of the Quality of Evidence

As we have noted, building evidence capacity within government involves increasing both the demand and supply for evidence to inform public deliberations, and this entails securing some level of agreement from the potential users as to when the evidence provided to them is good enough. Given the differences in criteria applied by evidence brokers, such as advocates for evidence-based policy, and the many clearinghouses discussed above, there are conflicting signals to policymakers. Researchers and evaluators need to be prepared to educate policymakers, senior government executives, program managers, and the public about the appropriate criteria for judging the quality of evidence as well as the value of using evidence to inform decision-making.

There are many sorts of claims based on evidence from research and evaluation that are pertinent to policymakers. Sometimes simple estimates of the occurrence of infections of a disease such as COVID-19 are needed, and other times the impact of a specific intervention on a specific subgroup of community members is requested. As shown in Table 2.3, the level of challenges to ascertaining the credibility of claims varies.

TABLE 2.3 ● The Array of Evidence Claims

Type of Claim	Data Required	Typical Data Analyses Employed	Criteria Used to Judge the Strength of the Claim	Typical Challenges	Examples
An estimate of the presence or incidence of a condition	Empirical data collected with a systematic tool or rubric from a population or relevant sample from that population	Simple counts or indices calculated according to clear rules	For quantitative data: measurement validity and measurement reliability For qualitative data: authenticity and auditability	• Incomplete coverage of population • Unrepresentative samples • Social apprehension bias in reporting • Inconsistent inputting of data across reporting entities	• Body mass index • Self efficacy • Chronic absenteeism • Infant mortality rate • Unemployment rate
A normative judgement about an incidence estimate	In addition to above, criteria from laws, regulations and/or experts to make judgements such as compliance, timeliness, adequacy	Comparison of the average of the incidence estimate to an established value	In addition to above, representativeness and size of the sample	In addition to above, lack of agreed upon criteria and/or cut offs to make normative judgements	• Compliance of regulated firms with air quality standards • Compliance of automobiles with mileage standards • Timeliness of handling social security claims

(Continued)

TABLE 2.3 ● *(Continued)*

Type of Claim	Data Required	Typical Data Analyses Employed	Criteria Used to Judge the Strength of the Claim	Typical Challenges	Examples
A relationship between two conditions	Empirical data on both conditions in the relevant time period	Correlations run between the two variables, e.g., Pearson's r	In addition to above, relevant correlations for benchmarking, e.g., previously run on the same jurisdiction or for other comparable jurisdictions in same time frame, and statistical probability estimates for sample values	In addition to above, incomplete contextual knowledge of other factors that may affect both conditions, e.g., spuriousness, or the relationship between the two, e.g., moderators	• Relationship between regular cardio exercise and heart disease • Lead in air and incidence of asthma • Relationship between driver's text messaging and automobile accidents
An estimate of the impact of one manipulated factor on an outcome	Empirical data on the intended outcomes both with and without the manipulated factor ("cause")	Random control evaluation; pretest with posttest design; difference-in difference design; propensity scoring; regression discontinuity design	In addition to above, knowledge about implementation of the alleged "cause," including contextual factors that may affect or moderate the causal relationship	In addition to above, refusals of participants; attrition of observed units from the study; and inadequate time frame for observation of the causal relationship	• Polio vaccine affecting contraction of polio • The use of nets over beds affecting contraction of malaria • Cash transfers to mothers affecting school attendance of children

An estimate of the impact of multiple factors on an outcome	Empirical data on the intended outcome and the various factors presumed to affect the outcome with all measured in the appropriate time frame	Various form of multiple or logistic regressions; Qualitative comparative analysis (QCA); Bayesian analyses	In addition to above, knowledge about the relationships among the multiple causal factors	In addition to above, failure to include potential causal factors in the analysis, and violation of assumptions to be met in order to use a specific analytical technique	• Behavioral interventions along with appropriate drug dosages affecting life expectancy after contracting the HIV virus
A prediction of a value for an outcome in the future (based on existing data)	In addition to above, an adequate and representative time period of data for all variables	Various form of multiple or logistic regressions; QCA; Bayesian analyses	In addition to above, estimates of the prevalence of conditions that affect the outcome into the future	In addition to above, inaccurate assumptions about the prevalence and levels of presumed causal factors and emerging contextual conditions into the future	• Enactment of sexual harassment policies regarding reporting and punishment of perpetrators predicting subsequent incidence of sexual harassment • Lowering of interest rates by a central bank predicting the magnitude of bank loans

The key in assessing the credibility of claims and the evidence supporting them is to recognize the relevant criteria to apply. Being clear and transparent about measurement accuracy and processes is always essential. Providers need to document and clarify the evidence trail of measurement regardless of what types of data are generated, or collected and analyzed. And then, depending on the sort of claim pertinent to the decision-makers, other criteria may also be applicable, as shown in Table 2.3.

While the term "validity" is sometimes used to describe evidence as if it is a quality that is either present or absent, that is misleading. There are multiple dimensions of validity and reliability that merit consideration, as discussed in this chapter. Importantly, while multicultural validity is pertinent to virtually any public policy consideration, it is not often discussed. In fact, steps taken to ensure cultural competence in evaluation work should always be discussed to enable the potential users to better judge the relevance of the evidence they receive.

There are partners out there for educating users of evidence, such as the federal agencies that have been leaders in establishing evaluation standards, such as the Departments of Education and Labor, and the many foundations that are promoting the use of evidence by government policymakers, e.g., Pew Charitable Trusts and the Gates Foundation. But since the signaling on the relative importance of different criteria is not always consistent, those doing evaluation work and those brokering the work to support managers and leaders in government need to be educated. They also need to be prepared to explain the nuances and potential limitations affecting the quality of evidence in an audience-friendly but complete manner.

Appendix 2.2 provides a checklist for users to assess the quality of evaluation reports and research studies.

Conclusion

In this chapter, we discussed how to assess the quality of evidence, and described widely accepted criteria for judging the quality of evidence, as well as evaluation and research study findings. We also illuminated some of the differences existing across diverse providers and synthesizers of evaluations and research regarding what constitutes sufficiently rigorous evidence. We then offered guidance on educating relevant stakeholders about the quality of evidence used for different purposes. In the next chapter, we expand upon the value and practical uses of evaluative thinking for measurement and evaluation work in government.

Exercises

1. You are presenting the findings and recommendations of a study to top leadership that was based on anonymous surveys undertaken on a large army base about the impact of sexual harassment training regarding reporting and punishment of perpetrators on the subsequent incidence of sexual harassment on the base 12 months after the training. List the questions you would raise to the study's authors regarding potential limitations to the study's findings before you brief the leadership.
2. Pick two of the websites listed below and identify and compare them on:
   a. the criteria they employ to assess the rigor of the evidence provided in the studies that they post and
   b. their ease of use for practitioners.

   https://ies.ed.gov/ncee/wwc/

   https://campbellcollaboration.org/

   https://clear.dol.gov/

   https://www.cebc4cw.org/

Resources for Additional Learning

Hart, Nick, and Meron Yohannes, eds. 2019. *Evidence Works: Cases Where Evidence Meaningfully Informed Policy*. Washington, DC: Bipartisan Policy Center. https://papers.ssrn.com/sol3/papers.cfm?abstract_id53766880.

Appendix 2.1

Limitations to the Credibility of Claims

Measurement Accuracy		
Limitation	**Potential Causes/ Defined**	**Examples**
Inappropriate Operationalization	Evaluators have insufficient knowledge about the concept of interest or the target population with which the concept will be measured, or the concept is impossible or too expensive to measure directly so approximate or "proxy measures" are used.	Questions for some psychological concepts (e.g., self-esteem, alienation) are standard; for other concepts (e.g., legal quality, sexual harassment), the means of operationalizing the occurrence of the concepts are still being explored. And questions or indices that have been validated for majority groups, or only men, may not be applicable to participants from different socioeconomic, racial, ethnic, or cultural groups
Purposeful Misrepresentation	Respondent intentionally distorts facts to protect themselves or hide a problem.	An agency official or program participant provides an answer that is technically accurate but is misleading as to the essence of the inquiry.
Accidental Misrepresentation	Faulty memory or records are not updated in a timely manner. Accidental misrepresentation is especially a problem when significant calendar time has elapsed.	An agency official or program participant unintentionally gives false information due to faulty memory of facts or events. Computerized inventory records may not be

Measurement Accuracy		
Limitation	**Potential Causes/ Defined**	**Examples**
		updated in a timely manner, creating a misleading impression of amounts in storage.
Social Desirability/ Evaluation Apprehension	The respondent tells the interviewer what he or she believes the interviewer wants to hear with the aim of receiving approval or a desire to please. Self-reporting may be more uncomfortable or not even acceptable for participants from different socioeconomic, racial, or cultural groups	Agency officials report that financial records accurately reflect inventory. New immigrants and non-citizens are hesitant to provide confidential information about themselves or their families in the Census.
Sleeper or Lag Effects	Effects lag beyond the time of measurement. In other words, what is being measured may be right, but the measurement is being taken at the wrong time.	The effects of television viewing on children's attitudes may not be immediate, but may be long-term. Other examples are business cycles, cycles in unemployment rates, or participation in welfare programs.
Change in Definitions	Redefining the data describing or monitoring an entity makes data from two or more time periods not comparable.	There may be changes over time or over jurisdictions in what is considered a "family" for qualifying for welfare, or what is considered a "misdemeanor" versus a "felony."
Lack of Dosage Differentiation	Measuring a treatment as simply received or not received when in fact program participants receive widely varying amounts of "treatment" (i.e., program services or policy) due to groupings,	Assuming that persons enrolled in a program receive the same amount of services, students in class receive the same amount of training, or taxpayers receive the same level of scrutiny.

(Continued)

(Continued)

Measurement Accuracy		
Limitation	**Potential Causes/ Defined**	**Examples**
	geographic areas, individuals, etc. Another type of treatment distortion is introduced when survey recipients give inaccurate information about their level of participation in programs or the benefits they receive.	
Mono-Operation Bias	Any one operationalization of a construct may underrepresent the construct of interest or measure irrelevant constructs, complicating inference.	Measuring attainment of a job as the only measure of the effectiveness of a job training program.
Mono-Method Bias	When only one method is used to operationalize the concept (e.g., self-report).	Using only body mass index to measure obesity; or using only self-reports on amount of time spent studying.
Lack of Cultural Insight	Failure to operationalize any phenomena, or interpret findings from measurement without taking the perspectives of the program participants into account.	Writing survey or interview questions without the input of representatives of the participants to be surveyed or interviewed, or interpreting responses without including representatives of the participants to interpret findings.

Measurement Processes		
Threats	**Potential Causes/Defined**	**Examples**
Lost in Translation	Questions are translated into multiple languages but the words do not really capture the same concepts.	Questions that include words such as political, bureaucratic, and "in compliance with the regulation" are not easily transferred into multiple languages.
Culturally Insensitive/ Intrusive Measurement Procedures	Failure to take the perspectives of the program participants into account when designing processes for observing or recording anything.	Failing to use representatives from the racial and ethnic groups who comprise focus groups of participants.
Multiple Judgment Calls	Questions rely too heavily on subjective assessments and different respondents may view the adjectives differently.	Questions that ask respondents to make distinctions between adjectives that may be interpreted differently, such as "poor, fair, average, and above average" may elicit different responses.
Capacity-Dependent Collection/Coding	Entering data from multiple locations may be overly dependent upon the capacity of those responsible for collecting and/or coding the data to carefully apply the same criteria in their decisions on how to collect or code; and high turnover, heavy workloads, and/or lack of technical capacity may render the collection/coding inconsistent across locations.	Busy front-line social service delivery staff, e.g., social workers, may not have the time to enter data; and staff in developing countries may not have the time nor technological support to input the data.
Insufficiently Prepared Data Collection and/or Coding (Intercoder Reliability)	Insufficient training of data collectors, interviewers, observers, and/or coders may render collection and/or coding inconsistent.	Overly ambitious timelines may push collection into the field too quickly, or efforts to save resources by cutting training may leave staff unprepared to ensure consistent collection and/or coding.

Causal Claims and Generalizability		
Threats	**Potential Causes/ Defined**	**Examples**
History or Intervening Events	The observed effect is due *not* to the program or treatment but to some other event that has taken place. For example, while a program is operating, many events may intervene that could distort pre- and postmeasurements as they relate to the outcome being studied.	A dramatic increase in media coverage on HIV/ AIDS distorts the measurements about the effect of a school-based program.
Lack of Multicultural Analysis	The observed effect is due *not* to the program or treatment for all participants, and/or the magnitude of the effect varies across participants, especially for participants from marginalized groups.	Failure to differentiate effects along cultural differences among the program participants.
Maturation	The observed effect is due *not* to the program but to the respondents growing older, wise, stronger/ weaker, etc., over time.	Juveniles often outgrow delinquent behavior as they age, making it difficult to disentangle maturation effects from the effects of a new community program. As people age, their health problems may become more pronounced, leading to an underestimation of the actual success of an exercise program to increase mobility (i.e., they would have been even worse off without the exercise program).
Testing or the Learning Curve	The observed effect being due to taking a test or being observed/measured several times. In a pre- and posttest design, group members could have	Participants in a training program learned from taking the test rather than from the program.

Causal Claims and Generalizability		
Threats	**Potential Causes/ Defined**	**Examples**
	scored better in the postperiod because they were more familiar with the test or measurement process and test situation.	
Program Not Fully Implemented	If inadequate resources or other factors have led to implementation problems, it is premature to test for effects. Even when programs or interventions have been implemented as prescribed by law, it is still wise for evaluators to measure the extent to which program participants or service recipients actually received the benefit.	The training on how to implement a new curriculum may not have been easily accessible to all teachers, so they were unable to implement as intended, or hardware and software needed for implementation were not delivered in time to all service providers.
Regression to the Mean or Regression Artifacts	The observed effect is due to the selection of a sample on the basis of extremely high or extremely low scores of some variable of interest. Change in the scores or values on the criterion of interest may be due to a natural tendency for extremely high or extremely low performers to fall back toward the average value. It would be misleading to attribute this change to the intervention. These threats arise when a program or other intervention occurs at or near a crisis point. To the degree that the fluctuation is random or occurrence idiosyncratic due to some cause of short duration, it is easy to incorrectly estimate to effects of	Participant exam scores, crime rates, and claims processing rates are all likely to rise and fall over time.

(Continued)

(Continued)

Causal Claims and Generalizability		
Threats	**Potential Causes/ Defined**	**Examples**
	whatever action or response is made.	
Selection or Selection Bias	The observed effect is due to preexisting differences between the types of individuals in the study and comparison groups rather than to the treatment or program experience. When the assignment of subjects to comparison and treatment groups is not random, the groups may differ in the variable being measured. "Volunteerism" can have a significant effect of its own.	Those who volunteer for a health promotion program may already be different (healthier) than those who do not.
Experimental Mortality	Individuals drop out of an experimental or treatment group between the pretest and the posttest, potentially exaggerating the magnitude of the observed effect because subjects who drop out of a program may have characteristics that differ from those who remain. Therefore, before-and-after comparisons may not be valid.	More highly motivated teens remain in a program designed to increase the teens' self-esteem.
Selection-Maturation Interaction	Selection biases result in differential rates of "maturation" or autonomous change within the treatment group. There may also be an interaction between selection biases and any of the other threats.	Volunteers for a job training program may be more disposed to follow the advice offered them.

Causal Claims and Generalizability		
Threats	**Potential Causes/ Defined**	**Examples**
Measurement Effects	A pretest or the process of taking observations may have a systematic effect on respondents, thus making the results obtained for a pretested or observed population unrepresentative of the unpretested universe.	Training participants who have taken a pretest may be sensitive to the intent of the training and pay more attention to the information highlighted by the test.
Situational Effects (Hawthorne, Staff, Novelty)	The observed effect is due to multiple factors associated with the experiment or study itself, such as the extent to which people are aware they are part of a study (Hawthorne effect), the newness of a program, and the particular time period in which a study takes place. This threat also includes atypical situation effects that make the selected context nonrepresentative on some dimension.	Instructors who volunteer to offer new training on sexual harassment in an agency or teachers who volunteer to implement an innovative teaching approach may be unusually enthusiastic due to the unique and timely nature of the topic.
Compensatory Equalization	When a treatment provides desirable goods or services, administrators or staff may provide compensatory goods or services to those not receiving treatment.	Teachers who are not implementing a new math curriculum (i.e., they teach a control or comparison group) work harder with the students.
Resentful Demoralization	Participants not receiving a desirable treatment may be so resentful or demoralized that they may respond more negatively than otherwise.	Comparison or control group members seek out training or treatment from other sources.
Treatment Diffusion	Participants may receive services from a condition to which they were not assigned, or learn from participants in the treatment group.	Students from new math curriculum treatment and comparison groups study math together outside of school.

(Continued)

(Continued)

Causal Claims and Generalizability		
Threats	**Potential Causes/ Defined**	**Examples**
Ambiguous Temporal Precedence	Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.	Schools in high-income areas may adopt healthy food policies (e.g., removing soda machines) in response to parent demands (as the parents already are pushing healthy eating at home).

Additional Threats to Generalizability		
Threats	**Potential Causes/ Defined**	**Examples**
General Selection Effects	Program results may only be applicable to the population/context that is directly studied, and may be more likely when evaluating or studying nonrepresentative cases, situations, or people.	
	Selection by Excellence	We may observe a situation because we believe it provides the best chance of seeing a hypothesized effect (e.g., the Job Corps increases the probability that teenagers will obtain jobs). However, a sound estimate of effect for an excellent program in one city may not be replicable in other locations. Thus, we may have only a "best practice" estimate.
	Selection by Expedience	We may observe a situation because it is accessible (e.g., available travel funds, proximity, persons who are willing

Additional Threats to Generalizability		
Threats	**Potential Causes/ Defined**	**Examples**
		to be interviewed). This is often a dangerous practice in that we have no way of knowing how representative the results are.
	Selection by Problem Severity	We may choose to look at locations or programs because we have some reason to believe that there is a severe problem there; e.g., we have some reason to believe that there is a contamination problem at a particular nuclear weapons production plant.
	Selection by "Where the Ducks Are"	We may observe locations or programs because they correspond to where large amount of dollars spent or large amounts of people are served. In this case, we are balancing limited resources, maximum payoff, and representativeness. Again, we need to be careful not to generalize to the universe of locations and programs, but it may not matter much to us if our chosen locations/groups account for a very large proportion (70%) of all dollars or activities.
Time Effects	The time frame of our observations may affect our estimates of important values; when using secondary data from other researchers, the data may	The performance of a weapons system tested during the day may bear no relationship to its performance at night.

(Continued)

(Continued)

Additional Threats to Generalizability		
Threats	**Potential Causes/ Defined**	**Examples**
	be so outdated that they are no longer relevant to the problem. Thus, although we may have a sound evaluation of some past regulation, policy, or program, there is no reason to believe that it bears any relationship to what is going on currently or future estimates.	
Geographic Effects	The evaluation may have been conducted in a specific area of the country or type of environment and its results are not generalizable to other settings.	A drug intervention program for urban youth in Chicago may not provide guidance on what should be done in rural Alabama.
Multiple Treatment Interference Effect	A number of treatments or programs are jointly applied and the effects are confounded and not representative of the effects of a separate application of any one treatment or program. Treatments are complex, and replications of them may fail to include those components actually responsible for the effects. An effect found with one treatment variation might not hold with other variations of that treatment, or when that treatment is combined with other treatments, or when only part of that treatment is used.	A drug abuse program designed for preteens may include several components (e.g., lectures, essay contests), making it difficult to separate out the effects of the different components.

Additional Threats to Generalizability		
Threats	**Potential Causes/ Defined**	**Examples**
Interaction of the Causal Relationship With Units	An effect found with certain kinds of units might not hold if other kinds of units had been studied.	When results are reported for schools rather than individual students, the inference may not hold at the individual level.
Interactions of the Causal Relationship With Settings	An effect found in one kind of setting may not be transferrable to other kinds of settings.	New health programs tried in Central America may not work in predominantly Muslim countries.
Context-Dependent Mediation	An explanatory mediator of a causal relationship in one context may not mediate in another context.	New health curricula may work in mixed sex schools but not in single sex schools (or vice versa).

Statistical Inferences		
Threats	**Potential Causes/ Defined**	**Examples**
Too Small a Sample Size	An effect or relationship of a specific size, regardless of the analytic approach used, is not statistically detected; there is low statistical power due to small sample size.	An effect of some magnitude in math achievement due to a new curriculum is not detected because too few students are included in the study.
Applying Statistical Analyses to Data Inappropriate for the Technique	The technique applied is not appropriate given the data and the underlying dynamics in measured relationships. Application of inappropriate statistical techniques for the data at hand may produce numbers that are misleading or incorrect. Each statistical technique is designed for application	T-tests should not be applied to ordinal measures (e.g., Likert 5-point scales), and nominal and short ordinal variables should be converted to dummy variables for use in regression.

(Continued)

(Continued)

Statistical Inferences		
Threats	**Potential Causes/ Defined**	**Examples**
	to certain types of data (i.e., nominal, ordinal, and interval/ratio), and for certain types of relationships between variables, e.g., linear.	
Violation of Assumptions Unique to a Statistical Technique	A particular type of test may not have sufficient power to detect an effect or relationship that is present, while another technique will be able to do so. Differences depend on assumptions made by the statistical techniques.	A t-test of means applied to two groups of respondents in which the variability is quite different may not provide an accurate test of differences. Regular OLS Regression should not be used to model nonlinear relationships.
Measurement Problems	If any of the variables used has a high degree of error, it threatens our ability to statistically identify relationships or differences and effects that are actually present; or other measurement problems such as unreliable proxy variables, or limited range in variables of interest.	If attitudinal scales contain adjectives that may have various connotations for responses (e.g., good, fair, outstanding), the responses may not be comparable across the sample; or if proxy measures are used and are inconsistently affected by other factors in the environment; or if age is assumed to be an important predictor, but in your sample the participants are only between 21 and 28 years of age.
Unreliability of Treatment Implementation	If a treatment that is intended to be implemented in a standardized manner is implemented only partially for some	When treatments or programs are implemented in a variety of contexts, the results may not be statistically

Statistical Inferences		
Threats	**Potential Causes/ Defined**	**Examples**
	respondents, effects may be underestimated compared with full implementation.	generalizable to all contexts.
Overfitting Models	Overfitting, that is including too many predictors to estimate an outcome of interest given the sample size.	When too many independent variables are included for relatively low sample size, the mathematical computations may result in showing inflated levels of both correlation and statistical significance, e.g., using 15 predictors in a regression using a sample of 50 units.
Specification Error	Specification effects may include either omission of other factors that may affect the outcomes of interest (similar to the history threat under internal validity) or inclusion of factors that are not relevant in an analytical model devised to predict specific outcomes.	When irrelevant variables are included in a regression model they may inflate the coefficient of determination (R^2), but not truly help predict the dependent variable of interest, and they may be collinear with predictors that are important and obscure their importance.

Appendix 2.2

A Checklist to Judge the Credibility of Evidence From Evaluation and Research Studies

Criteria	Data Collected via Quantitative Methods	Data Collected via Qualitative Methods
Role of Researcher(s)	• Cultural competence and humility • Reflexivity • No conflict of interest • Commitment to ethical practice and professional competencies	• Cultural competence and humility • Reflexivity • No conflict of interest • Commitment to ethical practice and professional competencies
Data Sources	• Reputable and unbiased sources of administrative data and other "big data"	• Relevant stakeholders included as participants
Sampling Approach/ Assignment	• Random or relevant and representative sampling of a large enough sample	• Relevant criteria for purposive selection of participants
Data Collection Techniques	• Objectively, clearly, and appropriately worded questions on surveys or collection tools for administrative data	• Appropriately focused and constructed observation, interviewing and focus group protocols
Timing of Measurement	• Appropriate for the questions addressed	• Appropriate for the questions addressed
Data Manipulation/ Coding	• Careful and appropriate handling of data collected, and of data collected by others	• Coding approach carefully and systematically conducted and reported
Data Analysis	• Techniques appropriate for the level of measurement of variables • Specification of model appropriate, and key assumptions met • Multicultural validity of findings checked	• Thematic analysis carefully and systematically conducted and reported • Multicultural validity of findings checked

Criteria	Data Collected via Quantitative Methods	Data Collected via Qualitative Methods
Validation Techniques	• Concurrent, content, and/or predictive validation of data and findings undertaken	• Member checking of findings with appropriate stakeholders
Recognition of and Addressing Limitations	Discussed/addressed limitations with: • Measurement validity • Measurement reliability • External validity • Internal validity (if causal claims made) • Statistical conclusion validity (if inferential statistics are used) • Multicultural validity	Discussed/addressed limitations with: • Authenticity of measurement • Auditability • Transferability and fittingness • Confirmability of claims • Multicultural validity