Evidence Reviews to Support Evidence-Based Policymaking

The **Foundations for Evidence-Based Policymaking Act of 2018** (the Evidence Act) directs Federal agencies to develop evidence to support policymaking. A crucial component of developing evidence is understanding what evidence already exists. This helps ensure that key learnings are incorporated into new and existing programming, and that the resources available for evidence-building activities are targeted towards areas where there are bigger evidence gaps.

With thousands of academic and policy articles and reports available, it can be difficult to know how to sift through it all. This guide provides an introduction to how to conduct a review of existing evidence in order to answer questions such as:

1. *Is there evidence that my program is effective? How much is known about whether and how my program works? What is still unknown?*
2. *If I am planning a new program, or making changes to my program, what practices or features should I consider incorporating? Are there similar programs or practices that have been shown to be effective?*

You can conduct an informal review to get an idea of what evidence exists, or you may want to do a more systematic evidence review. Depending on the level of detail or quality you need, you may have the expertise to conduct a review of evidence in-house, or you may need to consult with external experts. If you choose to conduct your review in-house, we hope this guidance will help you to identify the essential components of an effective review. If you plan to collaborate with experts, this guide is designed to prepare you to work with them efficiently and effectively by providing an orientation to key concepts.

For more in-depth treatment of some of the topics covered in this guide, we have included a list of resources at the end.

Key Concepts

Types of Evidence Reviews

While nearly any policy decision or program evaluation can benefit from conducting an evidence review, how **comprehensive** your review needs to be should be determined by scale and impact of your program. Expensive new initiatives, or initiatives with high policy importance, likely warrant a more thorough search for relevant evaluations than programs or program modifications that have lower stakes. While other types of evidence reviews exist, some key ones include:

A **literature review** is typically a more informal survey of existing work. It may simply provide an overview of (select) prior results, without attempting to be comprehensive or to systematically weigh the quality or findings of any given study.

A **systematic evidence review** attempts to find *all* published and unpublished evidence related to a specific research or policy question, using literature search methodologies designed to be transparent, unbiased, and reproducible. Systematic reviews will also often categorize the quality of the research and attempt to explain discrepancies in findings across research studies. It's important to note that, depending on the type of program you are running or proposing to run, there may be very specific criteria and protocols for conducting systematic evidence reviews. These can include requirements on the level of subject-matter and technical expertise needed to review the evidence. This is particularly true for evidence reviews required for evidence-based programs.[1] Before embarking on a review, make sure you understand what requirements apply to your situation.

Meta-analyses combine results from multiple quantitative studies, and use statistical techniques to synthesize and summarize results.[2]

Components of a Systematic Evidence Review

Some key components of a systematic evidence review are below. Again, note that, depending on your program, there may be specific requirements on the content of your review and the expertise needed to conduct it.

1. A description of the **selection criteria** that the researchers plan to use to choose the literature they include in their review. This could include:
    - The **specific question or outcome of interest** that the review is designed to address.
    - The databases they will use and the keywords they plan to search on in order to ensure that their review is comprehensive.
    - The criteria they will use to include or exclude identified studies from the review. For example, will they include only experimental or quasi-experimental studies, or will they also include qualitative evidence and formative or process evaluations? Will they only include literature published in peer-reviewed journals (the standard in academia), or will they also include reports from government or non-governmental agencies, or working papers and other unpublished materials?
2. A summary of the **results** of each study included in the review, including an explanation of which components of a given program were evaluated and may have been found to be more or less effective.

[1] See for example: Results for America. "Invest in What Works Fact Sheet: Evidence-Based Innovation Programs." Policy Report: October, 2015. http://results4america.org/tools/invest-works-fact-sheet-federal-evidence-based-innovation-programs/

[2] Source: Brown University Library. "What is a Systematic Review?" https://libguides.brown.edu/Reviews/types

3. An explanation, to the extent possible, of the **discrepancies** in results across different studies. For example, are there differences in the contexts (location, timing, population) in which the studies were conducted that could have impacted the results? What are the methodological differences (e.g. sample size, sample characteristics, presence and validity of a comparison group, outcome metrics and data sources) in the studies that make the results more or less credible and/or generalizable?
4. An explanation of any **critiques** that the researchers have of existing work, along with gaps in the evidence that they have identified.

Evidence Clearinghouses

Because assessing literature can be tricky, we recommend that you take advantage of the various **clearinghouses for evidence** that have been created. The staff at these clearinghouses have many years of experience in building evidence, so don't reinvent the wheel if you don't have to!

These clearinghouses typically fall into one of two kinds (although some have components of both). The first kind focuses on grading the quality of individual research papers or reports, often across a wide variety of topics. This kind of clearinghouse can help you find some of the better studies related to your work.

The second kind of clearinghouse typically focuses on grading the level of evidence supporting a specific intervention, by summarizing results across studies. Often these clearinghouses will also provide systematic reviews of evidence on certain topics. These clearinghouses tend to focus on specific policy areas, but if your work is related to the topics they cover, having access to existing systematic reviews can be extremely helpful.

A list of major clearinghouses is included at the end of this guide, along with some other publicly searchable evidence databases.

Meet Carlos…

Carlos is an IT manager at his agency. He has been charged with identifying potential ways of **reducing the number of people who fall for email phishing attempts** that can put the agency at risk of a cyber attack. Carlos and his team have brainstormed some different ways of addressing this issue, including:

- Training materials explaining what phishing is and how to avoid it
  - In-person
  - Online video or webinar
- Written materials
  - Emails
  - Manuals
- Running fake phishing attacks

Before Carlos decides which to invest in, he wants to know what has been most effective in the past, and whether there are promising practices that he can build into whichever program he decides on to make his program most successful.

The rest of this guide will follow Carlos as he gathers evidence to decide how to design his program.

Building Evidence to Support Programming Decisions

Whether you are conducting a literature review, building a systematic evidence review, or somewhere in between, here are some steps to take to incorporate evidence into your program or evaluation design.

Step 1: Assessing Relevance

While it may seem obvious, the first step in conducting a review is to filter the evidence based on how relevant existing studies are to your program.

- **Study Type:** *What* are you trying to learn from your review? Different types of studies can provide different types of evidence which may be more or less relevant, depending on your goals. For example, if your primary goal is to understand key learnings from how previous programs have been implemented, then reviewing process evaluations may be most useful, while if you want to know what programs have been most effective, then it is more important to look at impact evaluations. In areas where there has been little research, descriptive analysis of survey data may be all that is available.

This type of evaluation...	*Answers this type of question*[3]
Descriptive Studies	What is known about the landscape in which the program operates? I.e., who are the program participants, what are the program components,, what trends or patterns exist in data on participants or potential participants?
Formative Evaluations	Is the program, as planned, appropriate and feasible?
Process or Implementation Evaluations	Does the program operate the way it was intended to operate?
Outcome Evaluations	Is the program *associated* with positive outcomes for participants?
Impact Evaluations	Does the program *cause* positive outcomes for participants?
Economic Evaluations	Is the program cost-effective?

- **Intervention:** What are the *essential ingredients* or key characteristics of the program or intervention that was evaluated? In general, evidence is more relevant if it comes from an intervention more similar to your own. However, even interventions that may not be directly related to your program can offer insights. For example, simple reminders have been found to be effective at everything from increasing savings rates to getting people to enroll in health benefits. So think outside the box when you are considering what kinds of studies to look at, and then think carefully about what features of a program or its context might make it more or less relevant. In some evaluations you may find a discussion of "core components" of an intervention that drive its effectiveness.

- **Population:** *Who* does the program or intervention target? For example, an educational intervention that works well for children may not work for adults.

- **Context:** *Where* and *when* was the program implemented? For example, an intervention that works well in-person may not translate well to a virtual environment. Similarly, an intervention that works during an economic boom might work less well when the economy is struggling.

[3] Source: OMB Evidence Team and Office of Evaluation Sciences. 2020. *Evaluation 101*. Evidence and Evaluation Training Series. https://community.max.gov/x/yrHGe.

Carlos starts out by casting a wide net. He has friends in IT departments at other agencies who share some reports with him on how their training programs have worked in the past. But he also spends time searching through evidence clearinghouses for literature on different kinds of teaching methodologies, to see if there is anything he can learn about the best ways of teaching people, particularly adults. Lastly, he knows a professor from a local university who shares some academic papers with him, which he otherwise wouldn't have had access to because they are behind a paywall.

Carlos can't find any papers that specifically evaluate training programs designed to improve IT security. But there are a number of studies that compare online and in-person education. Most of these seem less relevant, since they are typically teaching a full semester's worth of material, while Carlos wants to provide much more limited information. He does find one study that looks at one-off, hour-long training sessions. It's a medical training, but he thinks it might still provide some insights.

Carlos is also interested in literature that looks at how the timing of when important information is provided might change how people respond to that information. He thinks that might be useful if he and his team decide to send emails with information about phishing.

Step 2: Assessing Quality

Determining the quality of an evaluation is not about assessing whether the results are positive or negative (or null), but rather *how well* the research supports those results.

- Is the chosen **research methodology** appropriate given the program being evaluated and the research question being asked?

  For example, process or implementation evaluations may lend themselves more to case studies, while for impact evaluations, the best available evidence comes from a randomized controlled trial.

- Are the methods **clear** and conducted with an appropriate level of **rigor**?

  The definition of rigor will depend on the type of evaluation being conducted. Unfortunately, understanding how rigorous a study is — and therefore how reliable its results are — requires expertise in the methods being used, and is outside of the scope of this guide. At the end of this guide, we have included links to resources that cover this in more depth.

- Do the **data** used accurately reflect the things they purport to measure? Are data collection methods and analyses clear?

  Sometimes there is no way of directly measuring key outcomes, so researchers use proxy data. For example, a study looking at whether new online materials are effective in helping people to apply for a government program might measure how frequently people click on

links to those materials. Depending on where the links appear and how they are advertised, this measure might or might not reflect how effective the materials are.

All studies should clearly document: (1) where their data come from, (2) how the data were collected, (3) who is represented in the data (for example, the full population of interest or only a select sub-population), (4) how any measures used are defined, (5) any important decisions that were made in preparing the data for analysis, and (6) what analytical methods were used.

- Does the study report **caveats** or **limitations**? Are these addressed appropriately? Do they seem reasonable?

  For example, even if the researchers were interested in program impacts, it is not always possible to conduct a randomized evaluation or a quasi-experimental evaluation. In this case, the researchers should be clear that, even if they have some evidence that the program is working, they cannot draw strong conclusions about whether the program caused the observed outcomes. Additionally, studies may be limited in their generalizability — for example, if the results depend strongly on the study sample or the context.

- Are the **conclusions** of the study reasonable based on the methods and the results?

  No study is perfect, and no study is going to be able to answer every question that researchers have about a program. Researchers should be clear about what their study can say about the program, where there is suggestive evidence but nothing firm, and what they still don't know.

For the two areas of programming he has decided to focus on, incorporating technology into training materials and sending informational emails, Carlos is primarily interested in learning about what models have been most effective. So he focuses on reviewing impact and outcome evaluations.

He first reviews ten different studies comparing in-person learning to hybrid online and in-person, and fully online, models. Of the ten studies, three are randomized evaluations that meet very high standards of quality. Five are quasi-experimental designs, but the authors have done a good job of testing the assumptions needed in order to make causal claims about the impact of the programs they are evaluating. Carlos discards the last two studies, because they make claims about the impacts of the programs that are not supported by the evidence presented.

He next turns to the literature on timing of informational interventions. These are almost exclusively randomized experiments and seem to be well-conducted. There is variation in the outcome measures, though, with some studies reporting on whether people clicked to open an email containing information, while others report on longer-term outcomes, like whether people followed through on the information provided. Carlos knows that opening and reading emails is important, but he is more interested in studies that show whether the emails actually changed behavior.

Step 3: Weighing the Evidence

Once you've determined the level of quality of the available studies, it may be possible to see whether the bulk of evidence supports the effectiveness of the intervention. Low-quality studies should be given little or no weight. But an intervention with several high-quality studies showing a positive effect is more likely to actually be effective than an intervention with several medium-quality studies showing a positive effect but only one high-quality study, which shows no effect. When weighing evidence, it's also important to take into consideration differences in the contexts or populations being studied to understand why different studies may have come to different conclusions, and what that says about the likelihood of success in your context.

In some cases, evidence might be mixed or inconclusive. In these cases, expert judgment might be needed to help draw actionable conclusions. Such cases might also highlight the need for additional evaluation, so may present an opportunity for future learning.

Null Results

A **null** result simply means that the study found that the program had no detectable impact. When weighing evidence, it's especially important to scrutinize null results. Depending on how a study was designed and implemented, a null result could be interpreted in two very different ways. One possibility is that a program or intervention did not work, but another possibility is that the evaluation wasn't sufficiently precise to detect its effectiveness. In colloquial terms, this is the distinction between 'evidence of absence' and 'absence of evidence.' In general, imprecise null results should receive little or no weight. It often requires expert consultation to determine what conclusion should be drawn from a null result.

The reports that Carlos got from other IT departments focus mainly on how easy it was to set up those programs, and whether users in the agency responded positively to them. This is valuable information for Carlos, but doesn't tell him whether the programs were effective in reducing the security risk.

Looking at online and in-person learning, one of the randomized evaluations found no differences in learning between students who participated in-person and those who participated in a hybrid model, while another study found positive effects. The randomized evaluation that had a context closest to what Carlos might want to do--a one-time session--found that online content that was more interactive was less effective than online content that included only text and graphics. All the studies identified gaps in understanding of which components of the courses were most important in improving learning.

Turning to the timing of email interventions, Carlos finds that there is some evidence that sending emails earlier in the week increases open rates, and also that providing information at key moments when people are about to make decisions may improve their likelihood to make use of that information. But he also learns that the right timing depends on the content of the message and the goal of providing the information.

Based on these insights, Carlos designs a series of online training modules. He also designs a series of emails with key takeaways from the trainings, which he hopes will remind people of what they learned and reinforce their knowledge. He also designs an evaluation of his new program, so that he can build on what he learned from reviewing the evidence and fill in some of the evidence gaps he identified.

Additional Resources

For Federal Executive Branch employees, the **Evidence and Evaluation Community MAX** page has an extensive list of helpful resources https://community.max.gov/x/iA_OJQ.

Evidence Clearinghouses and Research Libraries

Organizations that compile research on specific topics. Some also explicitly rate either the quality of the studies included, the efficacy of the programs under evaluation, or both. For a more comprehensive list, the Pew Charitable Trusts has compiled a list of research databases: https://www.pewtrusts.org/en/research-and-analysis/fact-sheets/2020/04/where-to-search-for-evidence-of-effective-programs

S=Rates study quality ***P=Rates program efficacy*** ***R=Posts systematic evidence reviews***

Clearinghouse Name	Topics Covered	Type
Clearinghouse for Labor Evaluation and Research (CLEAR) https://clear.dol.gov/	Apprenticeships, Behavioral insights, Career Academies, Child labor, Community college, Disability employment policy, Employer discrimination policy, Entrepreneurship, Job search, Literacy, Low-income adults, Mining, Older workers, Opportunities for youth, OSHA, Reemployment, Reentry, Veterans, Women in STEM	S, R
Results First Clearinghouse Database (note that this clearinghouse pulls data from nine other clearinghouses) https://www.pewtrusts.org/en/research-and-analysis/data-visualizations/2015/results-first-clearinghouse-database	Crime, Child and family well-being, Education, Employment and job training, Mental health, Public health, Sexual behavior and teen pregnancy, Substance use	P
What Works Clearinghouse https://ies.ed.gov/ncee/wwc/	Education: Literacy, Mathematics, Science, Behavior, Children with disabilities, English learners, Teacher excellence, Charter schools, Early childhood (Pre-K), K-12, Path to graduation, Postsecondary	S, P, R
Self-Sufficiency Research Clearinghouse https://selfsufficiencyresearch.org/	Asset-building, Tax policies and subsidies, Child care, Child support, Community development and housing, Education and training, Employment, Family formation, Food assistance, Income and poverty,	

	Health, TANF, Transportation	
IssueLab https://www.issuelab.org/	Consumer protection, Housing and homelessness, Energy and environment, Race and ethnicity, Government reform, Human rights, Animal welfare, Science, Substance abuse and recovery	
Clearinghouse for Military Family Readiness https://militaryfamilies.psu.edu/	Physical and behavioral health of military families	P
CrimeSolutions https://crimesolutions.ojp.gov/	Corrections and reentry, Courts, Crimes and crime prevention, Substance abuse, Juveniles, Law enforcement, Victims and victimization, Technology and forensics	P
Home Visiting Evidence of Effectiveness (HomVEE) https://homvee.acf.hhs.gov/	Home visiting programs for pregnant women and families with young children	P, R
Pathways to Work Evidence Clearinghouse https://pathwaystowork.acf.hhs.gov/	Employment for low-income individuals	S, P, R
The Campbell Collaboration https://campbellcollaboration.org/	Methods, Business and management, Crime and justice, Disability, Education, International development, Knowledge translation and implementation, Social welfare	R
Cochrane https://www.cochrane.org/	Health	P, R
Blueprints for Healthy Youth Development https://www.blueprintsprograms.org/	Youth development	P
Title IV-E Prevention Services Clearinghouse https://www.acf.hhs.gov/opre/research/project/title-iv-e-prevention-services-clearinghouse or https://preventionservices.abtsites.com/	Children and families: Mental health, Substance abuse, In-home parent skill-based, Kinship navigator	S, P
HHS Teen Pregnancy Prevention Evidence Review https://tppevidencereview.youth.gov/	Teen pregnancy and sexuality	S, P, R

Washington State Institute for Public Policy Benefit Cost Results https://www.wsipp.wa.gov/BenefitCost	Juvenile justice, Adult criminal justice, Child welfare, Pre-K to 12 education, Children's mental health, Health care, Substance use disorders, Adult mental health, Public health & prevention, Workforce development, Higher education	P, R

Further Resources on Conducting Evidence Reviews

- The Campbell Collaboration has a series of videos on evidence synthesis: https://campbellcollaboration.org/research-resources/training-courses.html
- Cochrane Training provides numerous resources on evidence syntheses related to health interventions: https://training.cochrane.org/
- Some common frameworks for structuring evidence reviews are covered here: https://guides.library.vcu.edu/health-sciences-lit-review/question
- Overseas Development Institute. *How to do a Rigorous, Evidence-Focused Literature Review in International Development.* *https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/8572.pdf*
- Innovations for Poverty Action. *Resources for Finding and Using Evidence Reviews and Evaluations.* https://www.poverty-action.org/publication/resources-finding-and-using-evidence-reviews-and-evaluations
- The various clearinghouses that conduct reviews (listed above) typically post detailed information about how they conduct their reviews, which can be helpful. Some recommended examples include:
    - Title IV-E Prevention Services Clearinghouse. *Handbook of Standards and Procedures.* *https://www.acf.hhs.gov/sites/default/files/opre/psc_handbook_v1_final_508_compliant.pdf*
    - Blueprints for Healthy Youth Development. *Blueprints Standards.* *https://www.blueprintsprograms.org/blueprints-standards/*
    - Home Visiting Evidence of Effectiveness (HomVEE). *Producing Study Ratings.* https://homvee.acf.hhs.gov/review-process/Producing%20Study%20Ratings
    - What Works Clearinghouse. *Standards and Procedures.* https://ies.ed.gov/ncee/wwc/Protocols#procedures
    - HHS Teen Pregnancy Prevention Evidence Review. *Review Process.* https://tppevidencereview.youth.gov/ReviewProtocol.aspx

Resources on Research Methodology and Rigor

There are many guides to evaluation available. Here we focus on some introductory guides to methodological rigor in different types of evaluations. For further resources, we recommend looking at the Evidence and Evaluation Community MAX page, linked above.

- Office of Management and Budget. Memorandum M-20-12.

https://www.whitehouse.gov/wp-content/uploads/2020/03/M-20-12.pdf (While not a guide, this offers useful information on federal evaluation standards and practices)

- Children's Bureau, Administration for Children and Families, U.S. Department of Health and Human Services. *Formative evaluation toolkit: A step-by-step guide and resources for evaluating program implementation and early outcomes.* *https://www.acf.hhs.gov/sites/default/files/cb/formative_evaluation_toolkit.pdf*
- Coalition for Evidence-Based Policy. *Which Comparison-Group ("Quasi-Experimental") Study Designs are Most Likely to Produce Valid Estimates of a Program's Impact? A Brief Overview and Sample Review Form.* http://coalition4evidence.org/wp-content/uploads/2014/01/Validity-of-comparison-group-designs-updated-January-2014.pdf
- Abdul Latif Jameel Poverty Action Lab. *Impact Evaluation Methods: What are They and What Assumptions Must Hold for Each to be Valid?* *https://www.povertyactionlab.org/sites/default/files/research-resources/2016.08.31-Impact-Evaluation-Methods.pdf*
- Mathematica Policy Research. *Understanding Types of Evidence: A Guide for Educators.* *https://www.mathematica.org/our-publications-and-findings/publications/understanding-types-of-evidence-a-guide-for-educators*
- Her Majesty's Treasury. *The Magenta Book, Annex A: Analytical Methods for Use with an Evaluation.* https://www.gov.uk/government/publications/the-magenta-book
- Alliance for Useful Evidence, Nesta. *The Experimenter's Inventory: A catalog of experiments for decision-makers and professionals.* https://media.nesta.org.uk/documents/Experimenters_Inventory.pdf