What makes research adoptable at scale?

A landscape analysis of policy research

Max Spohn, Elizabeth Linos, Jessica Lasky-Fink & Yunus Berndt
The People Lab · July 2026

ABSTRACT

Policymakers who want to use evidence to inform their decisions must assess the credibility, effectiveness, and relevance of the evidence base. While evidence clearinghouses and other policy-oriented frameworks aim to simplify the task of synthesizing and translating evidence across contexts, they lack common criteria and often ignore contextual factors that may influence real-world decisions. We propose a three-step framework for identifying programs that are credible, effective and adoptable, combining traditional indicators of evidence quality with contextual information that policymakers may rely on when making policy recommendations. We then apply this framework to an LLM-assisted landscape analysis of nearly 16,000 published research documents from 27 evidence clearinghouses and research organizations. We find that only 14% of studies meet basic standards of credibility, and just 65 program-outcome pairs are considered effective given our criteria. We discuss the sensitivity of our results to alternative criteria for identifying "what works" and propose an adoptability score to identify programs that may be more directly implementable for policymakers.

Policymakers who want to use evidence to inform their decisions face a complex task. They must assess whether a body of evidence is sufficiently credible, strong, and relevant to justify action at scale. In practice, this means answering a set of difficult questions: How confident should we be in research findings? Which programs have been shown to actually work more than once? And are these programs applicable – and feasible – in a given policy context?

The tools available to support policymakers in answering these questions are limited. Evidence is often fragmented across sources, inconsistently reported, and difficult to compare or synthesize across studies. Even institutions designed to synthesize research into usable guidance – including evidence clearinghouses (e.g., CLEAR; CrimeSolutions; Economic Mobility Catalog; Social Programs that Work) and evidence quality ratings (e.g., GRADE Working Group, 2013; Ruggeri et al., 2020) – reach different conclusions about what counts as good or strong

evidence. Even more importantly, they typically ignore contextual information like costs, political context, or implementation considerations. As we show in a separate report, *What policymakers want: Documenting a mismatch in policymaker demand and research supply*, these dimensions can play a pivotal role in policymakers' decisions.

A review of 14 prominent evidence clearinghouses identifies 22 distinct criteria used to classify programs. Some clearinghouses rely on as few as two criteria, while others incorporate up to 17. Virtually all prioritize study design and other indicators of internal validity, and the majority also propose methods of aggregating findings across studies in order to draw conclusions about effects. However, both criteria and their operationalization vary significantly across clearinghouses. For example, while some clearinghouses only consider randomized controlled trials (RCTs) to be credible study designs, others include quasi-experimental studies. Furthermore, signals of program adoptability and feasibility are rarely considered. These differences are not merely technical – they shape which programs are labeled evidence-based, which are considered promising, and which are excluded altogether.

We build on these existing frameworks to propose a three-step approach for identifying programs that work *and* are adoptable. Our framework focuses specifically on evidence on "what works" – that is, research that estimates the causal impact or effectiveness of programs, interventions, or policies – and incorporates both traditional indicators of evidence quality, such as study design or statistical significance, as well as the contextual factors that policymakers rely on when deciding whether to adopt a program. We then apply this approach to a corpus of policy research – nearly 16,000 research documents from 27 evidence clearinghouses and policy research organizations – to assess not only the quality and strength of the evidence, but also what types of contextual information are reported, and what is missing.

FINDING 01

Only **14%** of nearly 16,000 research documents meet basic standards of credible causal evidence; 72% are merely suggestive.

FINDING 02

After manual validation, we identify just **65** program-outcome pairs (**55** unique programs) considered "effective" based on our criteria.

FINDING 03

Adoptability information is scarce — many of the most pivotal dimensions are rarely reported in published research.

FINDING 04

The available adoptability information increases by **50%** when pooling across studies instead of relying on individual papers.

A framework for identifying credible, effective, and adoptable programs

Step 1: Identifying credible evidence on "what works"

The first step is to distinguish credible evidence from research that cannot reliably evaluate whether a program, intervention, or policy worked. To do this, we use four criteria that are applied at the study level, based on the research design rather than the results themselves:

- Evaluation method: is the method suited to identify causal effects of programs?
- Identifying assumptions: are key assumptions likely supported?
- Statistical power: is the evaluation sufficiently powered?
- External validity: are the results likely valid in real-world policy contexts?

Table 1 shows how we operationalize these criteria to classify studies into three categories: credible, informative, or suggestive.

Table 1. **Identifying credible evidence**

CATEGORY	DESCRIPTION	
Credible	**DEFINITION**	Policymakers can be confident in the results of the study when determining whether the tested program works or not.
	CRITERIA	The study met all of the following: • **Evaluation method:** randomized controlled trial (RCT), difference in differences (DiD), regression discontinuity design (RDD), or instrumental variables (IV); • **Identifying assumptions:** all identifying assumptions (based on the evaluation method) are likely supported; • **Statistical power:** reported power ≥ 80%, a minimum detectable effect (MDE) ≤ reported effect size, or sample size of ≥ 250 per condition; • **External validity:** was conducted in a natural environment with a sample drawn from the population of interest and no concerning selection into the study.
Informative	**DEFINITION**	Policymakers cannot be fully confident in the results of this study to determine whether the tested program works or not; more research is needed to increase confidence in the findings.
	CRITERIA	The study met the following: • **Evaluation method:** RCT, DiD, RDD, or IV; And one or more of the following holds true: • **Identifying assumptions:** some identifying assumptions (based on the evaluation method) are likely violated; • **Statistical power:** reported power < 80%, MDE > reported effect size, sample size < 250 per condition, or no power, MDE, or sample size reported; • **External validity:** was conducted in an artificial environment, sample was not drawn from the population of interest, or concerning selection into the evaluation.
Suggestive	**DEFINITION**	Policymakers should not rely on the results of this study to determine whether the tested program works or not, though the research may contain other relevant insights.
	CRITERIA	The study met the following: • **Evaluation method:** any method other than RCT, DiD, RDD, or IV (e.g., focus groups, correlational studies)

Example application

Consider an evaluation of a mailer designed to increase the take-up of social benefits. If the study uses a randomized or strong quasi-experimental design, meets key assumptions (e.g., valid assignment, limited attrition), is adequately powered, and is conducted in a real-world setting with the target population, it would be classified as **credible**. If one or more of these conditions are only partially met – if, for instance, the study has low compliance, limited statistical power, or an artificial setting – the study would be considered **informative**, meaning it provides useful but incomplete evidence. If the study relies on methods not suited for causal inference (e.g., focus groups or purely observational correlations), then it would be classified as **suggestive**, regardless of other features.

Considerations

There is no standard or widely accepted definition of what constitutes credible evidence. In fact, definitions and operationalization vary widely across academic disciplines. The criteria proposed in Table 1 thus offer a starting point that should be tested and adapted to specific contexts, as needed. In particular, setting thresholds, especially for statistical power, is a relatively subjective exercise. While 80% power is a common benchmark, what constitutes

a sufficiently large sample size also depends on the policy context: a sample of 250 individuals may be too small for a low-cost mailer intervention with relatively modest expected effect sizes, but might be sufficient for a more intensive program such as Crisis Intervention Training for police officers.

Step 2: Identifying which programs work

The second step in our framework is to aggregate evidence across studies to determine whether a program works. We recommend doing this at the *program-by-outcome* level, recognizing that policymakers may care about different outcomes depending on their context. This approach allows us to identify both which programs are effective for a given outcome, and which outcomes a given program has been shown to affect.

We use two criteria to identify whether a program works:

1. Number of credible studies evaluating the program-outcome pair;
2. Direction of results, measured as the share of credible studies reporting effective, null, backfire, mixed, or inconclusive effects.

Table 2 shows how we apply these criteria to classify each program-outcome pair into one of five broad categories: effective, null, backfire, mixed, or inconclusive. Among program-outcome pairs with *inconclusive* evidence, it's further possible to distinguish between those that lean towards effective, null, backfire, or mixed effects. In this brief, however, we focus on program-outcome pairs with *conclusive* evidence.

Example application

Consider again a study evaluating a mailer designed to increase take-up of social benefits. In this case, the program would be the mailer intervention and the outcome would be take-up, defined as applications for a specific benefits program. If multiple, independent, and credible studies – as identified in Step 1 – evaluate this program–by-outcome pair, we can begin to draw some conclusions about its effectiveness.

- If more than half of all credible studies find that the mailer increases take-up, the program would be classified as **effective**.
- If more than half of all credible studies report null effects, the program would be classified as **null**.
- If more than half of all credible studies report that the mailer decreases take-up, the program would be classified as **backfire**.
- If there is no clear majority – for example, if studies report a mix of positive and null results – the program would be classified as **mixed**.

If there are not multiple credible studies evaluating this program-outcome pair, the results would be considered **inconclusive**.

Considerations

The thresholds we propose, such as requiring at least two *credible* studies and a majority of results pointing in the same direction, are relatively subjective. We encourage researchers and policymakers to test how different assumptions affect conclusions about "what works" in their contexts. For instance, which programs would be considered evidence-based if we only required a single credible study showing that a program works, or if we required that over 75% of credible studies showed consistent effects, rather than 50%?

Table 2. **Identifying programs that work — classification and criteria**

CATEGORY	DESCRIPTION	
Effective	DEFINITION	There are sufficient credible studies testing the effect of Program X on outcome Y and the majority find desired effects.
	CRITERIA	Reported in 2+ credible studies (Step 1), *and* more than 50% report desired effects.
Null	DEFINITION	There are sufficient credible studies testing the effect of Program X on outcome Y and the majority find null effects.
	CRITERIA	Reported in 2+ credible studies, *and* more than 50% report null effects.
Backfire	DEFINITION	There are sufficient credible studies testing the effect of Program X on outcome Y and the majority find backfire effects.
	CRITERIA	Reported in 2+ credible studies, *and* more than 50% report backfire effects.
Mixed	DEFINITION	There are sufficient credible studies, but there is no majority among studies suggesting desired, null, or backfire effects.
	CRITERIA	Reported in 2+ credible studies, and 50% or fewer report desired effects, 50% or fewer report null effects, and 50% or fewer report backfire effects.
Inconclusive	DEFINITION	There are too few credible studies of Program X on outcome Y to know if the program works.
	CRITERIA	Reported in fewer than 2 credible studies. *Note:* this could mean only 1 credible study, or multiple informative/suggestive studies.

Step 3: Identifying adoptable programs

The third and final step is to determine which programs that "work" could also be adopted in practice. Evidence of effectiveness alone is often insufficient for adoption at scale. As we show in *What policymakers want: Documenting a mismatch in policymaker demand and research supply*, policymakers place substantial weight on contextual information, such as costs, political feasibility, and implementation information, when deciding whether to recommend a program for implementation. However, adoptability is highly context-specific. For example, the same program costs may appear high to some policymakers but low to others. As such, we use a single criterion to assess program adoptability – **availability of contextual and feasibility information** – rather than defining fixed thresholds. Access to these dimensions of evidence may allow policymakers to understand how translatable findings are to their own context and which organizational barriers they may face when attempting to implement a program.

Table 3 shows the nine specific dimensions that we consider when determining whether a program is adoptable. For each dimension, we construct a score reflecting whether relevant information is available. We then aggregate these scores into a composite adoptability score, constructed as the average of all nine dimensions. The adoptability score thus ranges from 0 (no adoptability information provided) to 1 (full adoptability information provided). Importantly, the adoptability score includes studies that are not classified as credible (as identified in Step 1). In other words, even informative or suggestive studies are included when calculating the adoptability score, as they may include the implementation information necessary for adoption.

Table 3. **Identifying adoptable programs — criteria**

DIMENSION	DO POLICYMAKERS HAVE ACCESS TO INFORMATION ON…	SCORING
Costs	…costs / cost-benefit estimates?	Binary
Implementation information	…what's required to implement the program? (13 items: implementation steps, actors, administrative capacity, existing processes, technology, coordination, barriers, employee buy-in, local discretion, time horizon, monitoring and enforcement, scaling, other)	Scale (avg. of 13 binary items)
Political context	…the political context of the program? (8 items: elite support, public support, stakeholder support, media coverage, electoral incentives, political stability, policy ownership, other)	Scale (avg. of 8 binary items)
Legal context	…the legal context of the program? (8 items: legal authority, compliance with existing laws, data privacy, legal risk, institutional legal capacity, inter-jurisdictional issues, legal precedent, other)	Scale (avg. of 8 binary items)
Heterogeneous treatment effects	…subgroup differences? (9 items: gender, race, geography, income, education, prior policy uptake, vulnerability, political affiliation, other)	Scale (avg. of 9 binary items)
Long-term outcomes	…the long-term outcomes of the program?	Binary
Location	…the locations where the program was evaluated?	Binary
Year	…the years in which the program was evaluated?	Binary
Program design	…crucial program design information? (5 items: theory of change, policy background, program development, method of delivery, program duration)	Scale (avg. of 5 binary items)

Example application

Consider again a mailer designed to increase take-up of social benefits. One study includes information on the program's method of delivery, duration, and policy background – but not on how the program was developed or its theory of change. In this case, three out of five factors comprising the **program design** dimension were reported, so the program would receive a score of 0.6 on this dimension. Scores for **implementation information**, **political context**, **legal context**, and **heterogeneous treatment effects** are constructed in the same way.

Imagine there exists a second study evaluating a sufficiently similar mailer that also reports the costs of the intervention. When creating the adoptability score, we consider all information available across studies. Thus, in this example, the **cost** dimension would also receive a score of 1 (i.e., cost information is available).

The final adoptability score is then calculated as the average of the nine dimensions.

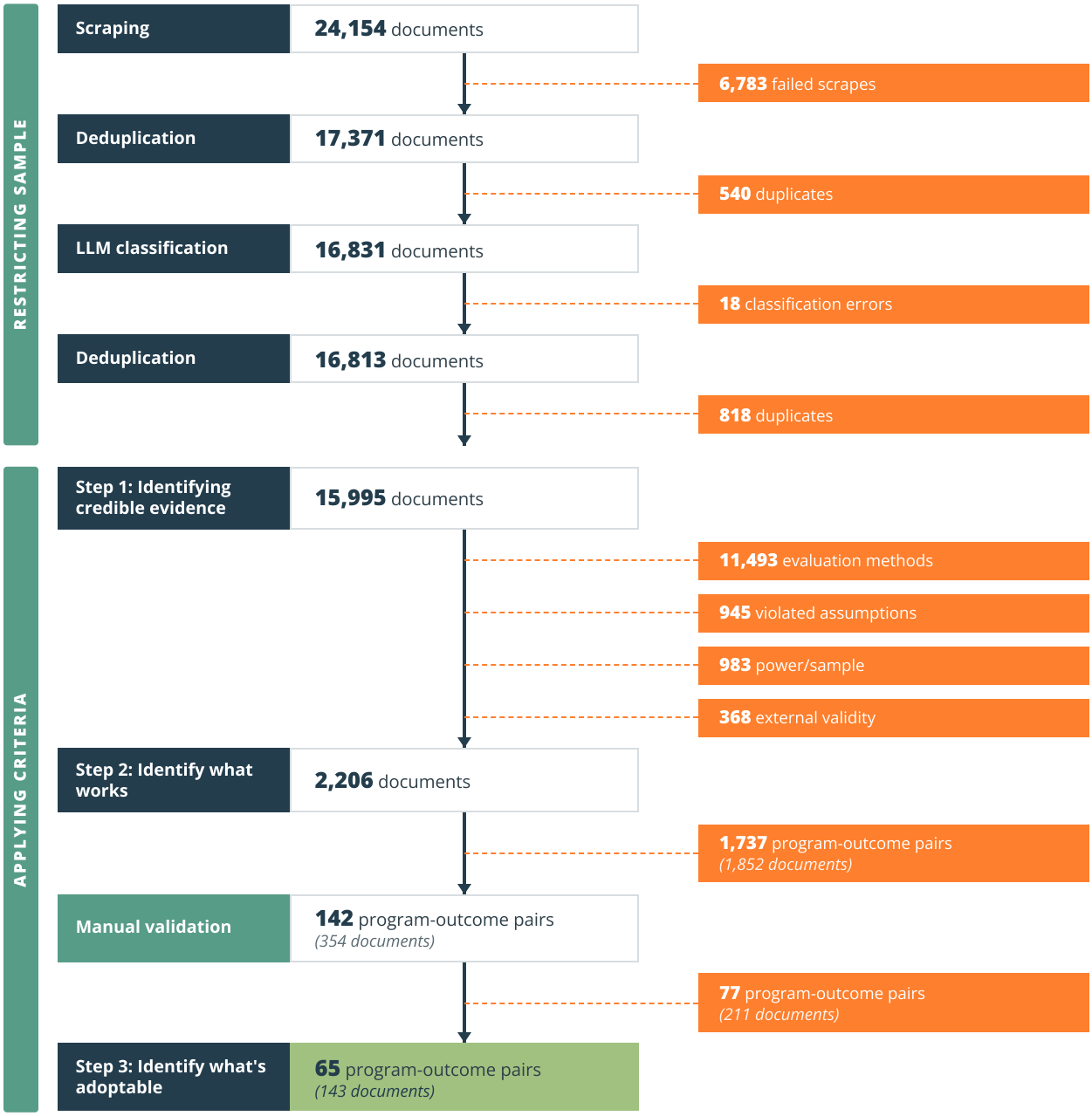
Considerations

While our adoptability score is calculated as an unweighted average of all nine dimensions, specific dimensions may be more important and relevant in some contexts. Thus, we encourage policymakers and researchers to consider weighting different dimensions (or factors) as needed to meet the needs of their context. Doing so could also lead to different conclusions about what constitutes adoptable evidence, which is worth exploring across policy contexts.

Application: A landscape analysis of policy research

To illustrate how this framework can be used in practice, we apply it to a large corpus of policy research to identify programs that are both supported by evidence and adoptable. We analyze 15,995 documents from 27 evidence clearinghouses and research organizations (e.g., CLEAR, Economic Mobility Catalog, BIT, OES, NBER Urban, etc.). We use GPT-4o and manual checks to classify each document along the evidence dimensions relevant to the proposed criteria. Figure 1 provides an overview of this process.

Figure 1. From 24,154 scraped documents to 65 program-outcome pairs that are both effective and adoptable.



Notes. The pipeline first restricts the sample (scraping, deduplication, LLM classification) and then applies the three-step framework. Orange boxes show documents or program-outcome pairs removed at each stage. The final adoptable set contains 65 program-outcome pairs (143 documents).

Step 1: How much credible evidence exists?

Applying our first step, we find that only a minority of research meets basic standards of credibility. Of 15,995 documents, roughly 14% are classified as **credible**, another 14% as **informative**, and 72% as **suggestive**.

We also investigate how sensitive this categorization of studies is to the specific criteria we proposed, as shown in Table 4. For example, removing internal validity criteria (identifying assumptions and statistical power or sample size cut-offs) would result in almost 23% of papers being classified as credible – and would, subsequently, lead significantly more program-outcome pairs to be classified as **effective** in Step 2. We also find that requiring that identifying assumptions be explicitly discussed would drastically decrease the number of papers classified as credible, potentially reflecting varying norms across fields or a propensity to only discuss assumptions when they are not obviously satisfied. On the other hand, the sample size criterion has a comparatively small impact on how many studies are classified as credible.

Table 4. Sensitivity of credible evidence classification to specific criteria

CRITERIA	% DOCUMENTS CLASSIFIED AS…			EFFECTIVE PROGRAMS
	CREDIBLE	INFORMATIVE	SUGGESTIVE	
Proposed classification criteria (Tables 1 & 2)	**14%**	**14%**	**72%**	**115**
Restrict credible studies to RCTs only	11%	13%	76%	104
No checks for identifying assumptions	16%	12%	72%	141
Require identifying assumptions to be discussed	2%	26%	72%	2
No checks for statistical power or sample size	18%	10%	72%	163
Require sample size of 100 per condition	15%	13%	72%	119
Require sample size of 1,000 per condition	10%	18%	72%	88
No checks for internal validity (assumptions + power/sample)	23%	5%	72%	223
No checks for external validity (environment, population, selection)	16%	12%	72%	130
No checks for internal or external validity	28%	0%	72%	295

Step 2: Which programs work?

Applying our second step, we group programs, treatments, and outcomes based on the semantic similarity of their descriptions, allowing us to aggregate results across studies even when the specific program or outcome names, as coded by large language models (LLMs), differ. To the extent possible, we try to avoid double-counting working papers and published studies, as well as various types of follow-up documents such as re-analyses of existing data and long-term follow-up studies.

Of the roughly 1,000 programs in our corpus of evidence that were tested in credible studies, we classify 115 as **effective**, 16 **null**, 1 **backfire**, and 30 **mixed** evidence program-outcome pairs when we group studies by their program descriptions and associated outcomes. Grouping studies by *treatment* descriptions (instead of program descriptions) and outcomes yields similar results, though specific programs and treatments do not always overlap. For example, while many common behavioral messaging strategies may be coded as separate treatments (e.g.,

planning prompts vs. social norms messages), they may be summarized as *Behavioral Messaging* on the program level. On the other hand, many post-welfare employment support programs may be comparable on the treatment level, but connected to specific welfare programs (e.g., MFIP vs. CalWORKS) on the program level. Combining a program- and treatment-level analysis allows us to identify a broader set of effective interventions.

14%
of studies are credible

65
effective program-outcome pairs

75%
of adoptability information is available

We again explore how our classification of "what works" varies by the criteria used (Table 5). We find that the classification of programs is relatively insensitive to the threshold of agreement between studies: requiring that more than 75% of available studies report the same directional effects – instead of more than 50% – only minimally reduces the number of effective program-outcome pairs we identify using our proposed criteria. However, the classification is dramatically affected by the requirement that there be at least two credible studies (i.e., replications) for a given program. Removing this requirement increases the number of program-outcome pairs considered to have conclusive results, and therefore the number that are considered effective, by a factor of 10. While replications are crucial for building confidence in a program's effectiveness (Maniadis et al., 2017), our results suggest that few interventions are systematically tested across multiple contexts and governments, highlighting a critical gap in academic incentives and policymaker goals.

Table 5. **Sensitivity of "what works" classification to different criteria**

CRITERIA	EFFECTIVE	NULL	BACKFIRE	MIXED
Proposed classification criteria (program level)	**115**	**16**	**1**	**30**
1 paper sufficient (program level)	1,421	296	80	68
75% inter-study agreement (program level)	109	10	1	42
Proposed classification criteria (treatment level)	**105**	**14**	**1**	**29**
1 paper sufficient (treatment level)	1,441	296	81	67
75% inter-study agreement (treatment level)	97	9	1	42

Following automated classification using GPT-4o, we manually investigated the resulting shortlist of effective programs. We validated whether programs, treatments, and outcomes were coded and grouped together correctly, whether results were correctly coded as effective, and whether we successfully removed follow-up studies and working papers as intended. Our final validated list of "what works" contains **65 program-outcome pairs** (55 unique programs), ranging from Summer Youth Employment Programs (e.g., Modestino, 2019) to Welfare-to-Work programs (e.g., Miller et al., 2000) and behavioral communications to nudge students to sign up for income-driven repayment plans (e.g., Office of Evaluation Sciences, 2016).

Step 3: Are these programs adoptable?

Finally, we apply our third step and examine which contextual and feasibility information is available in policy research documents. We find that the effective programs we identified in Step 2 receive, on average, an **adoptability score of 0.75**, suggesting a fair amount of adoptability information is available. As shown in Table 6, cost information is available for 95% of effective programs, long-term outcomes are available for 97% of programs, and location and implementation year are available for all effective programs. For the sub-dimensions that are scored on a continuous scale, we see that a fair amount of information on implementation and heterogeneity is available, while information on the political and legal context in which programs were evaluated is extremely scarce. In contrast, program design information is almost universally available.

In Table 6, we also show adoptability scores for papers. In other words, without aggregating across research documents that report results on the same program, how many individual documents report relevant adoptability information? Of note, we find that more adoptability information is available if we narrow the evidence landscape to credible studies and effective programs. This correlation between credible and effective evidence and the availability of adoptability information is primarily driven by dimensions that academics are increasingly paying attention to: cost-benefit estimates, heterogeneous treatment effects, and long-term outcomes. Yet, this obscures the fact that some of the dimensions policymakers may value the most, such as political or legal context, are still not widely available.

Table 6. **Adoptability scores and subscores for documents and programs**

DIMENSION	PAPER-LEVEL			PROGRAM-LEVEL	
	ALL N = 15,995	CREDIBLE N = 2,206	EFFECTIVE N = 354	ALL N = 4,147	EFFECTIVE N = 142
Overall	**0.44**	**0.49**	**0.52**	**0.54**	**0.75**
Costs	0.31	0.44	0.57	0.51	0.95
Implementation information	0.28	0.26	0.28	0.48	0.77
Political context	0.05	0.03	0.03	0.11	0.28
Legal context	0.05	0.02	0.02	0.10	0.16
Heterogeneous treatment effects	0.12	0.17	0.19	0.23	0.62
Long-term outcomes	0.39	0.58	0.62	0.57	0.97
Location	0.97	>0.99	1.00	0.99	1.00
Year	0.98	0.99	0.98	0.99	1.00
Program design	0.78	0.95	0.96	0.89	>0.99

Notes: Column 3 includes all papers that report effective results for the program-outcome pairs reported in column 5. Column 5 combines the program- and treatment-level aggregation of results and reports 142 program-outcome pairs (compared to 115 program-outcome pairs reported in tables above based on the program analysis alone).

On the one hand, these results are encouraging: a lot of effective programs are also adoptable in the sense that policymakers have access to necessary contextual and feasibility information. On the other hand, this exercise reveals an important challenge: the high adoptability scores are primarily achieved by grouping studies that evaluate the same programs to find and aggregate information associated with each of our adoptability criteria. To illustrate, the average adoptability score of credible studies identified in Step 1 – which are evaluated at the paper level – is merely 0.49. On average, studies report 44% of cost dimensions, 26% of implementation dimensions, and just 5% of political or legal considerations. Put differently, if policymakers were to solely search for individual studies, they would find relevant adoptability information in less than half of them. To find all relevant adoptability information, policymakers would need to find and synthesize multiple reports on the same program.

THE ADOPTABILITY GAP

On average, studies report 44% of cost dimensions, 26% of implementation dimensions, and just 5% of political or legal considerations.

Limitations

This work underscores both the promise and limitations of using LLMs for evidence extraction and synthesis. Traditional approaches for synthesizing large bodies of evidence – such as systematic reviews and evidence clearinghouses – require substantial resources and involve complex decisions about which studies to include and how to present findings (Pearson, 2024). Moreover, while these tools aim to reduce complexity, they often present large volumes of information that can be difficult to navigate and apply (Buckley et al., 2025). LLMs offer new opportunities to reduce the cost of synthesis by aggregating and summarizing large bodies of evidence, as demonstrated in this paper, but the exercise also revealed some key limitations.

First, using LLMs to extract and synthesize evidence requires first making a number of judgment calls around whether and how to categorize documents that report multiple studies, studies with multiple treatments and outcomes, and follow-up studies and working papers. While additional research could explore other ways of classifying research documents to ensure that all relevant information is captured, our results suggest that LLMs perform relatively well when extracting information from individual studies. However, aggregating and synthesizing results across individual studies is more challenging. In our landscape analysis, we relied on a state-of-the-art text embedding model to group programs and outcomes, but identified many instances of similar interventions that were treated as different programs, even though they should be treated as the same for the purpose of building an evidence base. For example, text-message reminders for vaccine appointments and text-message reminders for benefits interviews both reduce no-shows and likely operate through similar behavioral mechanisms. Yet, because they appear in different policy contexts and are described differently in the literature, the model does not consistently classify them as the same intervention or program. This limitation may reflect both the model choices we made, as well as the structure of the research literature and incentives within academic publishing. Future work could focus on improving approaches to addressing this challenge.

Second, our scraper designed to download individual documents from evidence clearinghouses appears to result in some policy areas and bodies of research being underrepresented in our landscape analysis. For example, among the 15,995 research documents we scraped, only 11 were published in *Justice Quarterly* and 14 in the *Journal of Experimental Criminology*, compared with 82 in the *American Economic Review* and 73 in the *Quarterly Journal of Economics*. This pattern primarily results from differences in our ability to successfully scrape documents from clearinghouses, though it could also be driven in part from our selection of clearinghouses, what studies clearinghouses decide to include, or underlying differences across disciplines in publishing norms and study designs

that may shape whether work is incorporated into clearinghouses in the first place. Overall, however, this suggests that relying solely on automated scraping of evidence clearinghouses to understand and synthesize existing evidence could skew the evidence base toward some disciplines while missing relevant work in others.

Despite these notable limitations, it is not obvious that humans are better suited to this task in practice. A panel of domain experts could potentially synthesize interventions across studies more accurately, but doing so at the scale required to build an evidence base would be both time and cost-prohibitive.

What's next

One of the main barriers to evidence use is knowing (and understanding) "what works." Finding, synthesizing, and translating evidence is a challenging task, and made more difficult by the fact that there is no common definition of or standard for what constitutes good evidence. The conceptual framework outlined here offers a starting point for assessing the credibility, effectiveness, and, importantly, adoptability of evidence. Applying this framework to a corpus of evidence scraped from 27 clearinghouses reveals some of the nuanced trade-offs and judgments inherent in evidence synthesis. For instance, how many replications of a program are necessary to allow us to confidently draw conclusions about a program's effectiveness? Does this (or should this) vary by context, including when the decision is politically risky or mistakes are costly? Should the availability of specific adoptability dimensions be weighted equally when considering what programs are adoptable or are certain types of information more important than others? We encourage researchers and practitioners alike to refine our proposed framework and criteria for different contexts.

While the advancing capabilities of LLMs will dramatically simplify the extraction and synthesis of evidence, these tools are still limited by what information is available. Importantly, our findings show a clear gap in the availability of adoptability information, such as legal and political context, that may be critical for policymakers and yet is rarely, if ever, included in research reports. Thus, developing a framework for assessing adoptability is only the first step to understanding what makes evidence adoptable at scale. Next, we must bring a rigorous academic lens to testing "what works" to make evidence more easily adoptable – from including additional implementation details in research reports to building sustained research-practitioner partnerships to support adoption of evidence-based solutions at scale.

ABOUT TPL

The People Lab aims to empower the public sector by producing cutting-edge research on the people of government and the communities they serve. Using evidence from public management and insights from behavioral science, we study, design, and test strategies for solving urgent public sector challenges in three core areas: strengthening the government workforce; improving resident–government interactions; and reimagining the production and use of evidence.

Acknowledgements: This research was supported by Arnold Ventures.

The People Lab

HARVARD Kennedy School
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

CONNECT WITH US

peoplelab.hks.harvard.edu
LinkedIn
Bluesky

References

GRADE Working Group. (2013). GRADE handbook for grading quality of evidence and strength of recommendations. https://gdt.gradepro.org/app/handbook/handbook.html

Maniadis, Z., Tufano, F., & List, J. A. (2017). To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study. *Economic Journal, 127*(605), F209–F235. https://doi.org/10.1111/ecoj.12527

Miller, C., Knox, V., Gennetian, L. A., Dodoo, M., Hunter, J. A., & Redcross, C. (2000). Reforming welfare and rewarding work: Final report on the Minnesota Family Investment Program, Volume 1: Effects on adults. Manpower Demonstration Research Corporation. https://acf.gov/sites/default/files/documents/opre/mfip_vol1_adult.pdf

Modestino, A. S. (2019). How do summer youth employment programs improve criminal justice outcomes, and for whom? *Journal of Policy Analysis and Management, 38*(3), 600–628. https://doi.org/10.1002/pam.22138

Office of Evaluation Sciences (2016). Income-Driven Repayment: Targeted Messages. U.S. General Services Administration. https://oes.gsa.gov/assets/abstracts/1603-1-Income-Driven Repayment-Targeted-Messages.pdf

Pearson, H. (2024). Scientists are building giant 'evidence banks' to create policies that actually work. *Nature, 634,* 16–17. https://doi.org/10.1038/d41586-024-03100-2

Ruggeri, K., Linden, S. V. D., Wang, C., Papa, F., Afif, Z., Riesch, J., & Green, J. (2020). Standards for evidence in policy decision-making. *Nature Research Social and Behavioural Sciences.* https://communities.springernature.com/posts/standards-for-evidence-in-policy-decision-making