Unexpected and Null Results Can Help Build Federal Evaluation Plans and Learning Agendas

Evaluation is a critical part of evidence-based policymaking. When conducting an evaluation of an intervention or program, our goal is to answer a specific question or test a specific hypothesis. After we design or introduce an intervention, we often expect it to result in changing outcomes in a certain direction (increase or decrease), or at a certain size or magnitude (such as a 2 percentage point increase). Yet even well-planned studies do not always meet expectations in terms of effect direction or size.

Recent research shows that null results are more common than we think, and occur for a variety of reasons. We know about this trend because agencies and researchers are increasingly publishing and discussing their null results more publicly.[1] Leading federal agencies publish null results, including the Department of Labor's Chief Evaluation Office (CEO), the Department of Education's Institute for Education Sciences (IES) and the Administration for Children and Families' Office of Planning, Research and Evaluation (OPRE). By sharing both expected and unexpected results, we can learn about what programs work, what effect sizes are realistic, and improve Federal evaluations.

What is an unexpected or null result?

- An unexpected result is one that runs counter to what prior evidence or informed hypotheses would suggest. This could mean an evaluation that shows no evidence of impact (null), a smaller or larger impact than prior evidence would suggest, or one that points to inconsistent program implementation. Such results can update our knowledge about the scope for change in large-scale government services.

- A null result, one type of unexpected result in an experiment, is when there is no statistically significant difference in outcomes between conditions: an intervention and a control (no-intervention) group, or between groups receiving two different versions of an intervention. This does NOT mean that one can conclude that the intervention has *no effect*. Instead, the observed impact was not large enough to stand out against background variation in the evaluation.

Some misconceptions about unexpected or null results in Federal evaluation:

Misconception: *Null results are rare.*

Truth: Not all interventions will be effective. As noted above, more null results are being published than before for interventions in health and social sciences. One-third of completed evaluations by the Office of Evaluation Sciences (OES) with federal agency partners had null results. As agencies conduct more research, they are likely to encounter studies that do not show evidence of an

---

[1] Turner et al. 2008; Franco, Malhotra and Simonovits, 2014; Kaplan and Irvin, 2015; Featherstone, Coffman and Kessler, 2018.
https://blogs.worldbank.org/impactevaluations/how-publish-statistically-insignificant-results-economics
https://blogs.plos.org/everyone/2015/02/25/positively-negative-new-plos-one-collection-focusing-negative-null-inconclusive-results/

intervention's (program, pilot, program enhancement etc.) success. That is how we learn and improve.

Misconception: A null result means my program or intervention did not work.

Truth: A null result may occur because the evaluation (rather than the program) did not work, and it is important to distinguish which is the case. First, check the study design. If your sample size is too small to detect a meaningful effect, then you cannot conclude that the intervention does not work. Even when you have a sufficiently large sample size, and a clear outcome that relates to your intervention and is measured well, you can rule out that we do not see evidence of an intervention's effect as large as what was initially projected in this specific evaluation. Second, explore the intervention mechanisms and implementation. If both study and intervention design are strong, then a null result could indicate the program may not be effective.

Misconception: A null result cannot benefit my agency.

Truth: We learn as much from what does show impact as what does not. Consider that a policy-relevant effect size is the smallest change that your agency would need to answer an evaluation question or take action. This can help you think about what level of intervention intensity or frequency is necessary to achieve that effect size. If a low-intensity intervention such as one-time letter does not demonstrate an effect, testing a higher-intensity intervention may be a clear next step.

Misconception: A null result will prevent me from doing future testing.

Truth: If a high-quality evaluation shows no evidence of impact, it can spur a series of more informed future tests. Agencies can ask whether the program was implemented as planned, whether it had a large enough sample size to detect a realistic and meaningful effect, and whether the intervention could have reasonably shifted the outcome. If these criteria were not met, then an agency may need to revise the intervention and re-test.

Misconception: Something unanticipated always happens in evaluations and program implementation, so why bother testing?

Truth: Agencies can learn from surprises and investigate what conditions would be necessary to ensure consistent implementation, expected sample size, and other features that promote a fair test. The departures from plan may highlight opportunities for increased training or information, more efficient allocation of staff, or other program features essential to success. An evaluation often leads to detailed documentation and learning about program implementation and agency data, which can be highly informative.

Misconception: If we get unfavorable results, the program will be stopped or discontinued.

Truth: Sharing null results from evaluations can promote more thoughtful implementation and design that could enhance program outcomes in the future. Careful planning in advance of an evaluation, such as researching a realistic and meaningful effect size and planning actions based on a range of possible results, can ensure program teams and leadership are prepared for and have an action plan for various results. These plans can include piloting additional changes, further testing, different targeting, additional types of assessment, and many actions beyond stopping a program.