10 Things Your Null Result Might Mean

Abstract:

After the excitement and hard work of running a field experiment is over, it's not uncommon to hear policymakers and researchers express disappointment when their trial returns a null result.

This guide explains that a null result rarely means "the intervention didn't work," even though that's the shorthand many people use. Instead, a null result can reflect the myriad design choices that policy implementers and researchers make in the course of developing and testing an intervention. After all people tend to label hypothesis tests with high p-values as "null results", and hypothesis tests (as summarizers of information about design and data) can produce large p-values for a number of reasons. Policymakers can make better decisions about what to do with a null result when they understand how and why they got that result.

Imagine you lead the department of education for a state and are wondering about how to boost student attendance. You decide to consider an intervention that offers students counseling and support to help them address challenges related to school attendance.

Here are 10 things a null from a randomized trial of that intervention could mean:

1. **Your study sample includes people whose behavior could not be moved by the intervention.**

You ask schools to randomize students into an intervention or control (business-as-usual) group. Some students in both your intervention and control groups will **always** attend school, while some students will **rarely** attend school, regardless of what interventions are or are not offered to them. Your intervention's success depends on whether it shifts behavior among potential responders or compliers.

If the proportion of potential responders is too small, then it may be difficult to detect an effect. In addition, your intervention may need to be targeted and modified in some way to address the needs of potential responders

> *How can you tell if:* The proportion of potential responders may be too small? Take a look at the pre-intervention participation rate. If it is extremely low, does that rate reflect low demand or structural barriers that may limit the potential for response. The OES tends

to use a threshold of 5% for this assessment, but it may differ depending on your own circumstances. (less than 5%)? Is it so high that it tells us that most people who could respond have already done so (say 85% or higher)? Even if there is a large proportion of potential responders, is it lower when you consider existing barriers preventing students from the use of counseling services that your intervention is not addressing?

2. **Your measure is not validated or reliable: It varies too much and systematically across sites.**

As a leader of this state department of education, you want to measure the effectiveness of your intervention using survey data on student attitudes related to attendance. Turns out only some schools administer a new survey measuring student attitudes, and those with surveys changed the survey items so that there is not the same wording differs across surveys or schools. If you observe no statistically significant difference on a survey measure that is newly developed or used by only select schools, it may be difficult to know whether the intervention "has no effect" or whether the outcome is measuring something different in each school because of different wording.

You then decide to use administrative data from student records to measure whether students show up on time to school. But it turns out that schools used a generic definition of on-time such that almost every student looks like they arrive on-time. An outcome that does not have enough variation in it to detect an effect between intervention and control groups can be especially limiting if your intervention could have had different effects on different types of students but the outcome measure can't capture that.

How can you tell if: Your null result may have more to do with the outcome measure than the intervention? Check to see whether the outcome is (1) collected in the same way across your sites and (2)if it means the same thing to the participants as it means to you. In addition, check on any reporting bias and if your study participants or sites face any pressure from inside or outside of their organizations to report or answer in a particular way.

3. **Your outcome is not validated or reliable: It varies too little.**

Given the problems with the survey measure, you then decide to use administrative data from student records to measure whether students show up on time to school. But it turns out that schools used a generic definition of on-time such that almost every student looks like they

arrive on-time. An outcome that does not have enough variation in it to detect an effect between intervention and control groups can be especially limiting if your intervention could have had different effects on different types of students but the outcome measure can't capture that.

> How can you tell if: Your null result arises from outcomes that are impossible to move? Pressures to report a certain kind of outcome faced by people at your sites could again yield this kind of problem with outcome measurement. So, it is again worth investigating the meaning of the outcomes as reported by the sites from the perspective of those doing the reporting. This problem differs from the kind of ceiling and floor effects discussed later in this guide: it arises more from the strategic calculations of those producing administrative data and less from the natural behavior of those students whose behavior you are trying to change.

4. **Your intervention theory and approach is mismatched to the problem.**

You delivered a counseling intervention because you thought that students needed support to address challenges in their home life. But students who had the greatest needs didn't end up meeting with a counselor, in part because they did not trust adults at the school. The theory of change assumed that absenteeism was a function primarily of a student's personal decisions or family circumstances and that the offer of counseling without changes to school climate would be sufficient; it did not account appropriately for low levels of trust in teacher-student relationships. Therefore, this null effect does not suggest that counseling per se cannot boost attendance, but that counseling in the absence of other structural or policy changes or in the context of low-trust schools may not be sufficient.

> *How can you tell if:* Your null result has to do with your problem definition and its correspondence to your theory of change? List all potential barriers and consider how they connect. Does the intervention as designed address only one of those barriers, and if so, can it succeed without addressing others? Are there assumptions made about one source or one cause that may undermine the success of the intervention?

5. **Your intervention does not represent a large enough enhancement over usual services.**

In your position at the state department of education, you learn that students at the target schools already were receiving some counseling and support services. Even though the existing

services weren't systematic or sufficient to boost attendance, the general content and frequency of the counseling services of the new intervention wasn't that different--existing counseling was once a month and the intervention ended up with show-up rates that were about the same. So this null effect doesn't reflect that counseling has no effect, but rather that the version of counseling your intervention offered was not effective over and above existing counseling services.

> *How can you tell if:* the relative strength of your intervention was not sufficient to yield an effect? Take stock of the structure and content of existing services, and consider if the lack of response to those services indicates that the theory of change or approach needs to be revised. If the theory holds, use existing services as a benchmark and consider whether your proposed intervention needs to include something supplementary and/or something complementary.

6. **Your intervention strength or dosage is too low for the problem or outcome of interest.**

After talking to experts, you learn that counseling interventions can build trust, but usually require meetings that are more frequent and regular than your intervention offered to have the potential for an effect. Maybe your "dose" of services is too small.

> *How can you tell if:* You did not have a sufficient "dose"? Even if no existing services tackle your problem of interest, consider what is a minimum level, strength or dose that is both feasible to implement and could yield an effect. When asking sites what they are willing to take on, beware of defaulting to the lowest dose. The more complex the problem or outcome is to move, the stronger or more comprehensive the intervention may need to be.

7. **Your intervention format was not reliable.**

In the schools in your study, counseling interventions sometimes occurred in person, sometimes happened by text message, sometimes by phone. The unplanned variation in format can reflect a host of selection bias issues, such that you cannot disentangle whether counseling as a concept does not work or whether certain formats of outreach did not work.

> *How can you tell if:* An unreliable format is the reason for your null? Were you able to specify or standardize formats in a checklist? Could you leave enough discretion but still

incentivize fidelity? Pre-specifying what the intervention should look like can help staff and researchers monitor along the way and correct inconsistencies or deviations that may affect the results. If nothing was specified, then the lack of consistency may be part of the explanation.

8. **Your intervention and outcome measure are mismatched to your randomization design.**

You expected counseling to be more effective in schools with higher student-to-teacher ratios, but did not block randomization by class size. In addition, if the intervention was delivered with lower dosage, less reliable formats, or altered in some way, then it may no longer have the potential to be more effective for students in high-class size schools .

How can you tell if: Inappropriate match is the reason for your null? Consider whether treatment effects could vary, or service delivery might occur in a cluster, or intervention concepts could spillover, and to what extent your randomization design accounted for that.

9. **Your statistical power is insufficient to detect an effect for the *intervention as implemented*.**

This may sound obvious to people with experience testing interventions at scale. But researchers and policymakers can fall into two traps:

i) thinking that with sufficient numbers of subjects in a field experiment, one does not need a strong intervention and can simply accumulate additional sample size at the unit of randomization to detect an effect of a weak intervention (even though strong research design cannot compensate for a weak intervention in any easy or direct way), or

ii) thinking that the intervention will be implemented exactly as designed (even though implementation dosage and strength is often not delivered as designed nor to as many people as expected). In these cases, you may end up with an imprecisely estimated effect that with more sample size and all of the intervention design issues addressed above, could have shown a positive or negative effect.

iii) thinking that the only relevant test statistic for an experiment effect is a difference of means (even though we have long known that differences of means are valid but low powered test statistics when outcomes are not nicely Normal.)

How can you tell if: your null result arises mostly from low statistical power? Recall that statistical power depends on (a) effect size or intervention strength, (b) variability in outcomes, and (c) number of independent observations (often well measured with sample size). The previous discussions pointed out ways to learn whether an intervention you thought might be strong was weak, or whether an outcome that you thought might be clear could turn out to be very noisy. A formal power-analysis could also tell you that, given the variability in your outcome and the size of your effect, you would have needed a larger sample size to detect this effect reliably (imagine, for example you learned that if you had known about the variability in administration of the treatment or the variability in the outcome (let alone surprises with missing data), in advance your pre-field power analysis would have told you a different sample size.) A different test statistic can also change a null result into a positive result if, say, the effect is large but it is not an effect that shifts means as much as moves people who are extreme, or has the effect of making moderate students extreme. A classic example of this problem occurs with outcomes that have very long tails –- like money. A t-test might produce a p=.20 but a rank-based test might produce a p<.01 <https://oes.gsa.gov/projects/gsa-auctions/>. The t-test is using evidence of a shift in means to reflect on the null hypothesis of no effects. The rank-based test is merely asking whether the treatment group outcomes tend to be bigger than (or smaller than) the control group outcomes (whether or not they differ in means).

10. Your null needs to be published.

If you addressed all the issues above related to intervention design, sample size and research design, and have a precisely estimated, statistically significant null result, it's time to publish. Your colleagues, and other researchers need to learn from this finding, so don't keep it to yourself.

When you have a precise null, you do not have a gap in evidence --- you are generating evidence.

What can you do to convince editors and reviewers they should publish your null results?: This guide should help you reason about your null results and thus explain their importance. If other studies on your topic exist you can also contextualize your results, for example following some of the ideas from Abadie 2019 <https://economics.mit.edu/files/14851> (For an example see how Bhatti et al in their

study of a Danish governmental voter turnout intervention <http://www.kaspermhansen.eu/Work/WEP_2017.pdf> used previous work on face-to-face voter turnout (reported on as a meta-analysis here <http://www.kaspermhansen.eu/Work/british_2016.pdf>) to contextualize their own small effects.

DRAFT – DO NOT CIRCULATE