WHAT IS THE RISK OF AN UNDERPOWERED RANDOMIZED EVALUATION?

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
NORTH AMERICA

Randomized evaluations can provide credible, transparent, and easy-to-explain evidence of a program's impact. However, a randomized evaluation requires sufficient statistical power to yield meaningful results. Budgetary, program, and timing constraints may create pressure to conduct an "underpowered" evaluation—but the risks of doing so are substantial. An underpowered randomized evaluation may consume substantial time and monetary resources while providing little useful information, or worse, tarnish the reputation of a (potentially effective) program.

WHAT IS STATISTICAL POWER?

The power of an evaluation reflects how likely we are to detect any meaningful changes in an outcome of interest brought about by a successful program.

HOW DO I KNOW IF THE STUDY HAS ADEQUATE POWER?

Power is determined by a number of factors. One important factor in discussing power and sample size is to consider what the smallest amount of change to an outcome would be that would make a program worthwhile. That is, what size change would we need to see to consider a program effective? Based on the size of this effect, the outcome chosen, and characteristics of the study design, we can run "power calculations" to estimate the sample size that may be needed.

WITHOUT ADEQUATE POWER, AN EVALUATION MAY NOT TEACH US MUCH.

Say we are evaluating a healthcare program that would be profitable if it decreased hospitalizations by 10 percent, but we are only powered to detect changes in hospitalization of 20 percent or more. If hospitalizations decrease by a (statistically insignificant) 15 percent after the program for study participants, we cannot be sure that the program decreased hospitalizations by 10 percent, but we also cannot be sure that the program decreased hospitalizations by zero percent (i.e. had no impact); thus, the evaluation offers inconclusive evidence on whether or not the program should be continued.

FAILURE TO FIND A STATISTICALLY SIGNIFICANT EFFECT CAN BE MISINTERPRETED AS THE FAILURE OF THE PROGRAM, RATHER THAN THE FAILURE OF THE EVALUATION.

When a study with insufficient power is inconclusive, we say that we find no evidence of an effect, but this does not mean that we have found evidence of *no effect*. However, funders, media, and the general public can easily conflate "finding no evidence of an effect" with a "finding of no effect." As a result, inconclusive findings can damage the reputation of an organization or program nearly as much as conclusive findings of no effect.

IF WE ARE CONCERNED THAT OUR STUDY MAY HAVE INADEQUATE POWER, WHAT CAN WE DO (OTHER THAN SIMPLY ADDING MORE PEOPLE)?

Sample size is an important factor in determining the power of a study, and detecting small differences between treatment and control groups may call for especially large samples. However, sample size is not the only factor that affects the power of a study. A number of adjustments can be made to a study's design including: reducing the number of different treatment groups if there are groups receiving different variants of a program, randomizing subjects at a more granular level (e.g., the individual rather than the group level), increasing take-up rates, increasing the likelihood that participants stick to their original treatment assignments, and minimizing attrition from the study.