

# Powering an evaluation to detect effects

Avery Hennigar, Daniel Friend, Rebecca Piatt, and Angela Valdovinos D'Angelo

An important underlying principle when conducting statistical analyses for any evaluation is that no matter how much data an evaluator collects, they can never prove or disprove a hypothesis with absolute certainty. Instead, evaluations can present findings that can reflect the outcomes or other characteristics of a population or populations based on what was observed in the sample. It is always possible that an evaluator's conclusion could be wrong because sometimes, just by chance, a sample does not represent the population. Evaluators should carefully design and execute an evaluation that minimizes the risk of any errors that could interfere with their ability to draw accurate conclusions.

This brief first describes common errors in statistical testing, then recommends five steps to help evaluators conduct a power analysis and calculate effect sizes they can use throughout an evaluation.



## Understanding Type I and Type II errors

In statistical testing, there are two types of errors that could lead to incorrect conclusions:

- **Type I error.** A false positive, which happens when an evaluator concludes a relationship exists when it does not. For example, an evaluator finds that a Healthy Marriage and Relationship Education (HMRE) program improves relationship skills for youth, but in reality, it does not.
- **Type II error.** A false negative, which happens when an evaluator concludes a relationship does not exist when it does. For example, an evaluator finds that an HMRE program does not improve relationship skills for youth, but in reality, it does.

An evaluator can aim to minimize a Type I error by choosing a statistical significance level. Significance level or  $p$ -value represents the chance of committing a Type I error. For example, if the  $p$ -value an evaluator selects is 0.05, this means there is a 5 percent chance of committing a Type I error.

### Box 8.1. Key terms

**Effect size.** A quantitative indicator that measures how strongly the HMRE program is related to an outcome. For example, in a causal evaluation, it is the magnitude of difference between two groups. In a descriptive evaluation, this could be the magnitude of difference between two time points.

**Minimum detectable effect (MDE).** The smallest effect that is likely to produce an impact estimate at a given statistical significance level.

**Power analysis.** An assessment of a study's chance of detecting meaningful impacts or effects if they exist.

Sources: Bloom 1995; IES 2021.

The Administration for Children and Families (ACF) provides grants to fund healthy marriage and relationship education (HMRE) programs to strengthen and improve the quality of relationships. The programs offer a range of services from relationship education for high school students to marriage and relationship skills building for adult couples. Grant recipients may be funded to also conduct descriptive or impact evaluations of their funded programs. Independent local evaluators support grant recipients in conducting their local evaluations. This brief is part of a larger evaluation technical assistance (TA) toolkit developed by Mathematica to help HMRE local evaluators understand key program evaluation concepts, common evaluation challenges, and strategies to prevent or overcome challenges. The briefs are standalone documents that can be read in any order. The TA toolkit was developed with HMRE program staff, their local evaluators, and other partners in mind, but it is also relevant to other program areas and organizations.

To avoid a Type II error, an evaluator needs to ensure the evaluation is adequately powered to detect effects. Power is driven by two factors: (1) the estimated size or magnitude of the effect (also called effect size); and (2) the sample size (Box 8.1). Before beginning an evaluation, evaluators need to minimize the occurrence of a Type II error by determining what effect size to expect and computing the sample size that will give them enough power to detect that effect size.

Effect sizes help put statistically significant findings in context and make sense of them. Although a  $p$ -value can provide information about whether a difference exists above and beyond chance, it does not reveal the size of this difference (Sullivan and Feinn 2012). In addition, it is possible that even if a difference is significant, it might not be meaningful. Evaluators should consult previous literature in the field, program staff, and participants' feedback to determine if significant differences are practically or substantively meaningful.

Accurately estimating the effect size is a process that starts before any data are collected. Designing and executing a study with adequate power to detect a statistically significant effect (if one exists) take careful planning and monitoring.



### **Step 1. Identify outcomes of interest that closely align with the program's theory of change**

The first step in a power analysis is to determine the expected effect size. To do this, the evaluator must identify the outcome measures of interest. For an evaluation to be successful, there should be a clear, well-supported path between how the intervention is expected to achieve its results and in which aspects of a participant's life changes are likely to be evident. Selected outcome measures should reflect what the intervention is expected to influence (Coster 2013). A logic model or theory of change can illustrate these linkages, and the evaluator should consult it when selecting outcome measures.

To identify specific outcomes, evaluators can consult the literature as well as speak with program and evaluator staff to answer the following:

- What is the program trying to change?
- Does this measure align with the program's activities?
- How can this outcome be measured?
- Would assessing this outcome be meaningful to participants in the program?



### **Step 2. Determine the minimum detectable effect for each outcome of interest**

After identifying an evaluation's outcomes, evaluators should next identify the minimum detectable effect (MDE) for each outcome (Box 8.1). To do this, evaluators can consult existing research to determine what change is reasonable to expect for a given outcome. The MDE should align with what other HMRE or similar programs have found in previous studies with similar participants at similar time points. For example, if an evaluator is examining an HMRE program serving primarily Spanish-speaking couples with a curriculum that integrates HMRE and content on economic stability, the evaluator should identify research on a similar program serving Spanish-speaking couples to inform MDE estimates.

Evaluators may consider reviewing findings from earlier large HMRE evaluations, such as [Building Strong Families](#),<sup>1</sup> [Supporting Healthy Marriage](#),<sup>2</sup> [Parents and Children Together](#),<sup>3</sup> or [Strengthening Relationship Education and Marriage Services](#).<sup>4,5</sup> Evaluators can use the effect sizes these evaluations detected to determine if their own estimated MDE is realistic. Note, however, that these studies have large sample sizes, and therefore have the power to detect smaller effects than studies with smaller sample sizes.



### Step 3. Run a power analysis to determine the sample size necessary to detect the MDE

The next step is to conduct a power analysis for each of the selected outcomes (Box 8.1). A power analysis uses MDEs along with other factors to estimate the likely sample size needed to statistically detect a difference of the specified effect size. There are several key pieces of information evaluators need to conduct most power analyses:

- 1. Power and significance level.** Power level is similar to the significance level or  $p$ -value discussed earlier. The power level represents the chances of *not* committing a Type II error. The minimum power level threshold is typically 80 percent, indicating that an evaluation has an 80 percent chance of not committing a Type II error. Power and significance level have an inverse relationship—increasing the power level increases the probability of a Type I error. The same is true for significance level; setting a lower significance level decreases the probability of a Type I error, but increases the probability of a Type II error. As a result, there is always a trade-off in these decisions. An evaluator must weigh the severity of each error type to determine where to set each threshold.
- 2. MDE.** This is the smallest effect the evaluation is powered to detect. In social science research, Cohen's  $d$ , a standard effect size metric, is typically used for MDEs, but other metrics exist given the type of outcome (Table 8.1). MDEs are often classified as small ( $d = 0.2$ ), medium ( $d = 0.5$ ), or large ( $d \geq 0.8$ ) (Carson 2012). These broad categories can be used as a general guide, but MDEs may be smaller or larger depending on the context of the evaluation, such as the design of the study, the timing of follow-up measures, the characteristics of participants, and the quality of the outcome measures (Ferguson 2009; IES 2021; Sullivan and Feinn 2012). For example, an evaluator who seeks to measure outcomes at two years post-program completion might expect smaller effects, as treatment effects tend to diminish over time. Or an evaluator may be interested in an outcome that is hard to measure or has no reliability information. In this instance, the evaluator might also expect smaller effects due to the measurement quality.
- 3. Type of hypothesis test.** The type of statistical hypothesis testing can influence power. For example, a one-sided test will require a smaller sample to adequately power an evaluation than a two-sided test does. Evaluators typically use a two-sided test because they are interested in whether the program has an effect regardless of whether the average difference between the research groups is higher or lower.

<sup>1</sup> Office of Planning, Research, and Evaluation. "Building Strong Families, 2002-2013." Administration for Children and Families. <https://www.acf.hhs.gov/opre/project/building-strong-families-2002-2013>.

<sup>2</sup> Office of Planning, Research, and Evaluation. "Supporting Healthy Marriages, 2003-2014." Administration for Children and Families. <https://www.acf.hhs.gov/opre/project/supporting-healthy-marriages-2003-2014>.

<sup>3</sup> Office of Planning, Research, and Evaluation. "Parents and Children Together, PACT Evaluation, 2011-2020." Administration for Children and Families. <https://www.acf.hhs.gov/opre/project/parents-and-children-together-pact-evaluation>.

<sup>4</sup> Office of Planning, Research, and Evaluation. "Strengthening Relationship Education and Marriage Services (STREAMS), 2015-2022." Administration for Children and Families. <https://www.acf.hhs.gov/opre/project/opre/research/strengthening-relationship-education-and-marriage-services-streams>.

<sup>5</sup> Evaluators can also review the local evaluation final reports from the 2015–2020 cohort of Healthy Marriage and Responsible Fatherhood programs. <https://www.acf.hhs.gov/ofa/programs/healthy-marriage-responsible-fatherhood/data-reports>.

**Table 8.1. Description of effect size estimates**

Index	Description	General guidance for interpreting an effect size	Notes
<b>Between groups</b>			
Cohen's $d$	$d = M_1 - M_2 / s$ where $M_1 - M_2$ is the difference between the group means, and $s$ is the standard deviation of either group	Small: 0.2 Medium: 0.5 Large: 0.8 Very large: 1.3	A useful measure during the planning stage when conducting a power analysis
Odds ratio (OR)	$\frac{(\text{Group 1 odds of outcome})}{(\text{Group 2 odds of outcome})}$ If OR = 1, the odds of the outcome are equally likely in both groups	Small: 1.5 Medium: 2 Large: 3	Used for binary variables to compare the odds of an outcome occurring as a result of one intervention compared with another
Relative risk or risk ratio (RR)	$\frac{(\text{Group 1 probability of outcome})}{(\text{Group 2 probability of outcome})}$ If RR = 1, the outcome is equally probable in both groups	Small: 2 Medium: 3 Large: 4	Compares the probabilities of outcomes occurring from one intervention to another
<b>Measures of association</b>			
Pearson's $r$ correlation	Range from -1 to 1	Small: 0.2 Medium: 0.5 Large: 0.8	Measures the degree of linear relationship between two quantitative variables
$r^2$ coefficient of determination	Range from 0 to 1 Typically expressed as a percentage	Small: 0.04 Medium: 0.25 Large: 0.64	Represents the proportion of variance in one variable explained by the other

Note: Adapted from Sullivan and Feinn (2012).

There are many statistical programs that evaluators can use to conduct their power calculations, including free or open-source ones such as [G\\*Power](#). Some of these calculators have user-friendly interfaces that produce estimates of sample size based on the details the user enters about the evaluation, such as MDEs, type of statistical test, and desired significance level.

Power analyses are outcome specific, so evaluators should choose a sample size large enough to give the study power to detect all outcomes. For example, an evaluator might pick the largest sample size specified across all the outcomes to make sure the study is powered to detect all the outcomes' effects. It can also be a good idea to consider different scenarios and run a few calculations to understand what the study will be powered for if the sample size is not achieved (for example, assuming the program was only able to enroll enough people to achieve 75 percent of the intended sample size). This allows evaluators to identify which outcomes are well powered or underpowered if the desired sample size is lower than expected.



#### **Step 4. Review power and sample size estimates with program staff to determine monthly enrollment targets**

Once evaluators determine the sample size the evaluation needs, they can discuss these results with program staff and ask if these enrollment goals seem feasible. It can help to break down the desired sample size into monthly enrollment goals. When calculating monthly targets, evaluators

should be sure to consider times of year when enrollment might be slow (for example, in November and December around the holidays) and devise a plan to over-enroll in busier months to make up for it. If the monthly enrollment goals are not feasible as the evaluation progresses, evaluators may need to go back to Step 1 and revise their outcome measures to select an outcome that could be adequately powered if there were a more realistic estimate of sample size.

Running initial power calculations when designing a study is an important first step, but evaluators should remember they might have to adjust power calculations as the evaluation progresses. Evaluators should consider running these calculations on a regular schedule (at least annually) to determine how the pace of enrollment is aligning with power estimates.



## Step 5. Calculate and report the effect size estimate and compare it to the broader field

After completing follow-up data collection and conducting the planned evaluation analyses, a best practice is to calculate the actual effect size for the outcomes of interest. Table 8.1 provides an overview of common effect size estimates that evaluators can calculate. Evaluators should report the effect size estimates in their final report and other study publications, such as in the abstract and results sections. Effect size estimates are important to the field—for example, in conducting meta-analyses and to inform MDEs in future studies.

## References

- Bloom, H. S. “Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs.” *Evaluation Review*, vol. 19, no. 5, 1995, pp. 547–556.
- Coster, W. J. “Making the Best Match: Selecting Outcome Measures for Clinical Trials and Outcome Studies.” *American Journal of Occupational Therapy*, vol. 67, no. 2, pp. 162–170.
- Ferguson, C. J. “An Effect Size Primer: A Guide for Clinicians and Researchers.” *Professional Psychology: Research and Practice*, vol. 40, no. 5, 2009, pp. 532–538.
- Institute for Education Sciences. “Effect Size Basics: Understanding the Strength of a Program’s Impact.” 2021. [https://ies.ed.gov/ncee/edlabs/regions/west/relwestFiles/pdf/4-2-3-14\\_Effect\\_Size\\_Infographic\\_Final\\_508c.pdf](https://ies.ed.gov/ncee/edlabs/regions/west/relwestFiles/pdf/4-2-3-14_Effect_Size_Infographic_Final_508c.pdf).
- Office of Evaluation Sciences. “Effect Size and Evaluation: The Basics.” <https://oes.gsa.gov/assets/files/effect-size-evaluation-basics.pdf>.
- Sullivan, G. M., and R. Feinn. “Using Effect Size – Or Why the P Value Is Not Enough.” *Journal of Graduate Medical Education*, vol. 4, no. 3, 2012, pp. 279–282.

**Suggested citation:** Avery, Daniel Friend, Rebecca Piatt, and Angela Valdovinos D’Angelo (2024). “Powering and Evaluation to Detect Effects.” OPRE Report #2024-146. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

**Acknowledgements:** Many people contributed to this toolkit. First, we acknowledge staff at the OPRE in the Administration for Children and Families at the U.S. Department of Health and Human Services. We are particularly grateful for the direction and feedback from Samantha Illangasekare, Rebecca Hjelm, and Kathleen McCoy. We want to extend our gratitude to Sarah Avellar and Angela D’Angelo for reviewing drafts of the briefs. We also extend our appreciation to Effie Metropoulos and Bridget Gutierrez for editing and Yvonne Marki-Korosec and Gwyneth Olson for designing the graphics in this toolkit.