

Coming Soon!

Average effect sizes for outcomes are coming soon. The information below explains why effect sizes are important.

Why should I care about effect sizes?

As practitioners and policy makers, we want to know if a program will work for our population. In fact, program evaluations try to answer that very question: "Does this program result in favorable effects, and if so, are they big or small?" Knowing and understanding "effect sizes"—how big or small a program's impact is—can guide important decisions. It makes sense to use a program that has large, favorable effects on the problems we're trying to address rather than using a program that has small (or no) effects.

What is an effect size?

An effect size is a standardized measure that examines how much one variable (such as a program) affects another variable (such as an outcome of interest). NREPP examines effect sizes that reflect the size of an intervention's impact as measured by the difference between how much the intervention and comparison groups improved on outcomes of interest (such as substance use or mental health disorders).

NREPP effect sizes

Effect sizes may be measured, calculated, and reported in many different ways—for example, as an odds ratio, a Hedges' *g*, or a Cohen's *d*. NREPP calculates Hedges' *g* effect sizes for both continuous and dichotomous (yes/no) outcomes, because this measure takes small sample sizes into account. In addition, whenever possible, NREPP calculates an effect size that is adjusted for pretest differences.

Because an effect size actually provides an estimate of the impact, NREPP also looks at the confidence interval around an effect size. A confidence interval provides an indication of the likely range of an intervention's effect. Confidence intervals are defined by two numbers, one lower than the result found in the study and the other higher than the study's result.

What is the difference between effect size and statistical significance?

The phrase "statistical significance" is often used when discussing the impact of a program on an outcome of interest. But what does that mean? Statistical significance is the probability that the observed difference between two groups is not due to chance. An effect size may be found statistically significant (that is, likely not due to chance) or nonsignificant (that is, perhaps due to chance).

What is an average effect size?

Often, a study will examine the impact of a program on multiple outcomes, such as substance use and depression. It might also use multiple measures to examine the impact on one outcome—for instance, the researchers might have measured the severity of depression and the frequency of depression. To provide a better understanding of the overall impact of the program on depression, NREPP combines these two separate measures using a statistical method called meta-analysis. The meta-analysis yields an average effect size for the program impact on one

outcome (such as depression). NREPP also combines separate measures across studies, so study 1 might measure the frequency of depression, and study 2 might measure the severity of depression. These results can similarly be combined for an average effect size.

How can some outcomes be rated ineffective when they have bigger average effect sizes than other outcomes rated promising?

Like most statistics, average effect sizes give us valuable information, but there are limitations. For instance, meta-analysis (which is used to calculate NREPP's average effect sizes) takes a study's sample size into account, so studies with large sample sizes have more weight than studies with small sample sizes. But our average effect size does *not* take into account the strength of the studies. The ratings of effective, promising, and ineffective *do* take the studies' strength into account and therefore give us a different kind of information.

Let's assume we have a study with a large sample size that shows the program has a large positive impact on depression, but the study design is very weak. On the other hand, we have a study that has a moderate sample and shows essentially no program impact, but the study design is a high-quality randomized controlled trial and therefore very strong. The first study would have a greater impact on the average effect size through the meta-analysis (resulting in a large average effect size), while the second has a greater impact on the outcome rating (resulting in a rating of "ineffective").

How can I compare effect sizes?

Although effect sizes give us concrete numbers, interpreting the size of an effect is a difficult task. For example, the effect size of "Program Alpha" on depression could be presented as Hedges' $g = 0.3$. But is "0.3" big or small? It's hard to answer that question without some context. For instance, depression may be an outcome that can be influenced pretty easily with a good program or it may be an outcome that is really difficult to change, even if the program is excellent.

To help compare effect sizes across programs, we will present an *average effect size* on each outcome. More precisely, the effect sizes across all programs that address each outcome (such as depression) will be combined through meta-analysis (see above). The average effect size provides some context to help you understand if the effect of one program is bigger or smaller compared with other similar program effects.

For instance, if Program Alpha's effect on depression is " $g = .50$," you might conclude that the program is doing a pretty good job addressing depression, if the average effect " $g = .30$." On the other hand, if the average effect size is " $d = .70$," you might conclude that Program Alpha is not as strong as some other programs in addressing depression.

When will I be able to see the average effect size across outcomes?

Publishing an average effect size when there are only a few values for its basis is not very meaningful. That average can be very unstable—it can move up or down too easily when one value is added to the pool of existing values. When you have a large pool of values, however, adding another has a much less direct effect on the average.

Because we want to publish stable average effect sizes, we will offer this feature when a sufficient number of programs and outcomes have been reviewed. Stay tuned!