REL WEST
Regional Educational Laboratory At WestEd

Effect Size Basics:
Understanding the Strength of a Program's Impact

Purpose

This quick-reference guide provides basic information about the measure called "effect size" to inform the selection of evidence-based interventions. It provides the following information:

- What is an effect size?
- What is the difference between effect size and statistical significance?
- What influences the effect sizes reported in studies?
- How can effect size be interpreted to inform decisions?

Intended Use

Education research increasingly includes "effect sizes" in discussions of findings from studies of policies, practices, interventions, programs, services, and curricula.[1] Accordingly, it is helpful for educators to understand the definition of effect size, what factors influence reported effect sizes, and how effect sizes can be interpreted. This guide serves as a companion to the REL West guide, The Basics of Reviewing a Research Study, to assist state education agency staff, school district staff, and school staff as they review research to identify high-quality, evidence-based programs that meet their needs. This guide presents some basic concepts and considerations for reviewing research that reports effect sizes; however, it is not exhaustive. Additional resources are included at the end of this guide.

What is an effect size?

- Effect size is a measure of the strength or magnitude of the effect of a program on an outcome (or the strength or magnitude of the association between a program and an outcome) relative to a benchmark.[2]

1 This guide will use "program" as shorthand for policies, practices, interventions, programs, services, and curricula.

2 Effect size can also indicate the strength or magnitude of the association between a program and an outcome such as when effect sizes are reported in correlational or descriptive studies. In these cases, the "effect" in "effect size" does not necessarily mean the program caused the outcome.

- Effect sizes are most useful when they are from studies with target populations and outcome measures that are similar to the ones of interest.
- A common benchmark used in education studies is the standard deviation[3] of the outcome for the group that did not get the program (sometimes called the control or comparison group). For example, if a research article states, "The program had an effect size of 0.15 on the math assessment," this should be interpreted as the average math assessment score for the group that received the program was 15 percent of a standard deviation larger than the average math assessment score for the group that did not receive the program.
- Effect size units are "standardized" so that effect sizes from different studies can be compared to one another.
- Effect sizes can be positive or negative. A positive effect size is desired if the program aims to increase a desired outcome (for example, the program aims to increase reading proficiency). A negative effect size is desired if the purpose of the program is to decrease an unwanted outcome (for example, the program aims to decrease absences).

What is the difference between effect size and statistical significance?

Effect size and statistical significance each measure something different and both are important pieces of information to consider when using research to inform decisions about programs. This table compares effect size and statistical significance along several dimensions.[4]

Dimension	Effect Size	Statistical Significance
Definition	A measure of the strength or magnitude of the effect of a program on an outcome compared to a situation without the program.	The probability of observing a difference at least as large as the one between the two groups (for example, program versus no program) even if the true difference were zero.
Benchmark	Standard deviation	p-value

3 The standard deviation is a measure of how spread out the scores are around the average (or mean) on the outcome measure.

4 Many education researchers consider effect size an important piece of information when considering adoption of a program even though the Every Student Succeeds Act (ESSA) does not consider effect size as a factor for determining a program's tier of evidence.

Dimension	Effect Size	Statistical Significance
Use	To determine the strength or magnitude of a program's effect on the outcome.	To determine if observed differences between program and no-program groups (or between two groups receiving different programs) are statistically significant.
Key value threshold(s)	There is no single conventional threshold for determining a meaningful effect size.	The conventional threshold for a treatment effect being deemed "statistically significant" is $p < .05$.
Range of values	Infinite negative or positive values.	Positive values between 0 and 1.
Example	"The average assessment score was 56.34 for the students that received the math program while the average assessment score was 51.05 for the students that did not receive the math program. The math program had an effect size of 0.11." Meaning: The program increased math scores by 11% of a standard deviation.	"The average assessment score was 56.34 for the students that received the math program while the average assessment score was 51.05 for the students that did not receive the math program. This difference of 5.29 points was statistically significant at $p < .05$." Meaning: There was a 5% probability of observing a difference between the two groups at least as large as 5.29 points if the true difference between the two groups was zero points.

What influences the effect sizes reported in studies?

When reviewing a research study, it is important to remember that reported effect sizes are not only influenced by the effect of the program but also by characteristics of the study design and other factors. Below are several of many factors that can influence reported effect sizes.

Effect sizes reported in a research study tend to be smaller when….

- they are from the most rigorous research design, experimental studies (RCTs)
- there is less difference between what was implemented as part of the program versus no-program (or other-program) conditions
- the outcomes are difficult to measure
- the outcomes are measured long after the program was completed
- the sample represents the general population (for example, students of all socioeconomic backgrounds)

Effect sizes reported in a research study tend to be larger when….

- they are from less rigorous research designs such as quasi-experimental (QED) or descriptive studies
- there is more difference between what was implemented as part of the program versus no-program (or other-program) conditions
- the outcomes are easier to measure
- the outcomes are measured soon after the program was completed
- the sample represents a specific population (for example, students of lower economic backgrounds)

How can effect size be interpreted to inform decisions?

Policymakers and practitioners often want to know the "practical significance" of a program's effect size. Consider one or more of the following:

- **How does the effect size of a program compare with the effect sizes of other programs that focus on the same outcome?**

  One way to think about whether a reported effect size is meaningful is to compare it to the effect sizes of other programs that target the same outcome. This is particularly helpful when one has multiple programs to choose from for potential adoption. It requires comparing effect sizes across either individual research studies, or average effect sizes from research syntheses, or meta-analyses.

  > *For example, the Reading Wizard program for grade 4 students had an effect size of 0.26 on a standardized reading assessment. The Reading Buddies program for grade 4 students had an effect size of 0.15 on the same standardized reading assessment. So, Reading Wizard has a stronger effect than Reading Buddies.*

When comparing effect sizes across studies, syntheses, or meta-analyses, it is important to remember that the comparisons are relevant insofar as the target populations and outcome measures are similar to the ones of interest to the policymaker or practitioner.

- **How does the effect size of a program compare with a typical year of growth or change on the outcome of interest?**

  Another way to think about the meaningfulness of a program's effect size is to compare it to what would be expected in terms of growth or change on the outcome of interest in the absence of the program. This is a useful comparison when considering academic outcomes that are measured by standardized assessments such as reading, math, or science. Information on "typical" or "average" growth can often be found in the manuals for standardized tests.[5]

  > *For example, as students progress from grade 5 to 6, the average annual student reading growth, based on several nationally normed reading tests, is an effect size of 0.30.[6] Reading Wizard reports an effect size of 0.30 for grade 5, which mean it leads to no greater than average annual growth in reading in grade 5. Reading Buddies reports an effect size of 0.45 for grade 5, which means it has a greater effect than the average growth in reading for grade 5.*

  This way of interpreting effect size is not possible if there are no available data on typical growth. Also, existing data are often based on the general population of students, which may differ for a specific population of students (for example, English learner students).

- **How does the effect size of a program compare to the gap between different groups of students on the outcome of interest?**

  A third way to think about the meaningfulness of a program's effect size is to compare it to average gaps for demographic subgroups on the outcome of interest.[7]

  > *For example, the average White-Hispanic performance gap on grade 4 mean NAEP reading scores is 0.77, with Hispanic students, on average, scoring lower than White students. Reading Wizard reports an effect size of 0.35 for Hispanic students in grade 4. This means the average effect size of Reading Wizard for Hispanic students in grade 4 equals almost half of the White-Hispanic performance gap on the average grade 4 NAEP reading scores.*

  This way of interpreting effect size is not possible if there are no available data on the gaps. Also, data on existing gaps are often based on nationally representative samples which may differ from the local context.

5 Hill et al. (2007).

6 Scammacca, Fall, & Roberts (2015).

7 See Hill et al. (2007) for examples of these gaps.

Sources Cited and Additional Resources

Effect Size Basics

What Works Clearinghouse. (2020). *What Works Clearinghouse Procedures Handbook, Version 4.1.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/handbooks.

Ferguson, C. J. (2009). *An effect size primer: A guide for clinicians and researchers Professional Psychology: Research and Practice, 40*, 532–538. https://psycnet.apa.org/doi/10.1037/a0015808

The Difference Between Effect Size and Statistical Significance

Fan, X. (2010). Statistical significance and effect size in education research: Two sides of a coin. *The Journal of Educational Research, 94*, 275–282. https://doi.org/10.1080/00220670109598763

Greenland, S., Senn, S. J., Rotthman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *p* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*, 337–350. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877414/

Sullivan, G. M. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education, 4*, 279–282. https://doi.org/10.4300/JGME-D-12-00156

Wasserstein, R. L., & Lazar, N. A. (2016). *The ASA statement on p-values: Context, process, and purpose.* https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108

Factors Influencing Effect Sizes Reported in Research Studies

Bakker, A., Cai, J., English, L. *et al.* (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathemathics* 102, 1–8. https://link.springer.com/article/10.1007/s10649-019-09908-4

Interpreting Effect Sizes to Inform Decisions

Bloom H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289–328. https://www.mdrc.org/publication/performance-trajectories-and-performance-gaps-achievement-effect-size-benchmarks

Hill, C. J., Bloom, H. S., Black, A. S., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172–177. https://srcd.onlinelibrary.wiley.com/doi/10.1111/j.1750-8606.2008.00061.x

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncser/pubs/20133000/

Kraft, M. A. (2018, December). *Interpreting effect sizes of education interventions* (Ed Working Paper 19-10). Annenberg Institute at Brown University. https://files.eric.ed.gov/fulltext/ED602384.pdf

Scammacca, N. K., Fall, A., & Roberts, G. (2015). Benchmarks for expected annual academic growth for students in the bottom quartile of the normative distribution. *Journal of Research on Educational Efficacy, 8*, 366–379. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4696502/

IES Institute of Education Sciences

This product was prepared under Contract ED-IES-17-C-0012 by Regional Educational Laboratory West, administered by WestEd. The content does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.