Resource Guide: Effect Sizes

Purpose

The purpose of this EBITE Guide is to help educators understand and interpret the basics of effect sizes when reading reports or studies on the outcomes of educational interventions, programs, and practices. Knowing how to interpret and compare effect sizes is important when making a choice about a particular intervention or practice to adopt. Beyond this brief introduction, we provide some additional recommendations on resources and readings for those interested in more details or further information.

What are Effect Sizes?

Effect sizes are standardized measures that simply summarize the impact of an intervention on a specific learning outcome. Larger effect sizes indicate larger effects for the intervention. While simplistic in scope, they can be very complex to determine, and are typically provided in reports or articles that share the results of evaluation studies. To add to the complexity, there are many possible kinds of effect sizes. The general idea is to compare an average score on an educational outcome (such as an end-of-term math assessment score, or reading assessment) between a group of students that receives an intervention and a comparable group of students that does not. Through this comparison, an effect size can tell you how large or small the difference is between the two groups, relative to the natural variability among all the scores. The "standardization" refers to this amount of variability in the outcome scores, which is called the standard deviation. Our discussion below and resource section links will help make these ideas clear.

Effect Sizes and Level 4 Interventions

According to ESSA, level 4 interventions are defined as "*demonstrating a rationale based on high-quality research findings or positive evaluation that such activity, strategy or intervention is likely to improve student outcomes or other relevant outcomes*." Effect sizes are important ways of showing amount of improvement or change in an educational outcome. When making decisions about whether to use a particular intervention, the size of the effect must be large enough to justify the decision to adopt the intervention or program. But "large enough" is a value judgement, and depends on your needs and capacity for the intervention ultimately selected for implementation.

Note!
Effect sizes can be used to help build a rationale for a level 4 intervention – but other features such as cost, the school context and demographics of students in the original study, and ease of implementation are also important.

When a school chooses to put a particular intervention or program into practice, educators are anticipating that the effect for their own students, classrooms or schools will likely be the same or similar to what was reported in an evaluation study. However, school contexts or settings that differ from those of the original evaluation study and even slight variations in how an intervention is implemented can affect the size of the reported change. The quality of the evaluation study design can also affect the reported effect size. Weakly designed evaluation studies can yield unreliable estimates of an intervention's effect size. Thus, the effect size is only *one piece of information* that educators should consider before adopting a new educational intervention or practice.

Why are Effect Sizes Important?

Effect sizes are often compared between two different interventions designed to change the same achievement outcome. However, interventions vary in terms of time, cost, training, and materials. Comparing effect sizes from evaluation reports or publications on several different interventions for the same intended outcomes can help schools make good decisions while balancing issues such as cost and

other factors. Educational improvement and the size of potential change is important, but a school may opt for a simpler, low-cost intervention with modest expected effects over an intervention that is expensive and challenging to implement and monitor, even if the anticipated effect size may be larger.

One Size Fits All? A Cautionary Tale

Effect sizes don't just come in one…size. There are many different types of effect sizes that could be found through varying kinds of evaluations. It is not essential that educators know how to determine an effect size value, but it is important to feel comfortable with interpretation when reading the literature about potential effects of an intervention or practice, particularly when making a decision among two or more interventions.

One of the most common forms of effect size is called "*Cohen's D*," where D stands for "difference." For Cohen's D, interpretation of an effect size is fairly straightforward; distance is in standard deviation units. For example, if D = .50, this implies that the distance between means of the intervention and comparison groups is .50 standard deviations. Half of a standard deviation difference is considered a moderate or medium effect. We could also interpret this .50 effect size as indicating that the average score in the intervention group was 50 percent of a standard deviation larger than the average score in the comparison group. Another familiar effect size measure is the correlation coefficient between two variables, "*Pearson's r*" or just simply "r." Correlation assesses the strength of a linear relationship between two variables. A correlation of +1.0 is a perfect positive correlation; a correlation of .30 is considered moderate or medium effect.

Over many years of study on educational interventions, there are effect size values that have come to be considered as small, medium, and large effects (Table 1). However, we urge educators to pay close attention to all aspects of a study – such as its evaluation design quality, reliability of measures, cost, complexity, and student sample – when interpreting the value of an effect size and deciding to implement an intervention. It is also important to pay attention to the nature of the "business as usual" or practice being used in the comparison group. If the comparison program is different in evaluations of a targeted intervention, the effect size is likely to be different too! That is, the **meaningfulness** of a particular effect size value is a judgement that educators must make in balance with a constellation of many study features.

Table 1. Effect Size Conventions for Cohen's *D* and Pearson's *r*

	Effect Size		
Type	Small	Medium	Large
Cohen's D	.20	.50	.80
Pearson's r	.10	.30	.50

Summarizing Collections of Effect Sizes

Evaluations of the same practice/intervention in different schools or classrooms could be conducted in many different ways (for example, through a randomized trial or a correlational study, or they could rely on different outcome measures or analysis designs) and thus result in different effect sizes. Given school/classroom and evaluation design differences, it's challenging to summarize effect sizes into a single overall estimate of an intervention's effectiveness. Essentially, this is the work of research compendiums and evidence repositories such as the What Works Clearinghouse (WWC;

https://ies.ed.gov/ncee/wwc/), Evidence for ESSA (https://www.evidenceforessa.org/), or the Ohio Evidence-based Clearinghouse (https://essa.chrr.ohio-state.edu).

These Clearinghouses provide the most up-to-date and summary information on tested interventions. Many of these compendiums, the WWC in particular, use sophisticated meta-analysis methods to summarize effects across multiple evaluation studies of the same intervention. The resulting effect size summaries are viewed as the most trustworthy in representing the expected change for students that can be attributed to an intervention. However, it's important for educators to be mindful of the many factors that can affect the size of an effect – such as cost, implementation complexity, or quality of evaluation design.

Visible Learning

You may be familiar with John Hattie's work on *Visible Learning* (https://visible-learning.org/). Hattie[1] (2009) reviewed many different meta-analyses on hundreds of educational practices and interventions. A primary goal of Hattie's work is to place achievement-related effect sizes along a single continuum in order to identify those educational activities/interventions associated with greatest change. Based on this work, Hattie identified effect sizes ≥ .40 as those in the "zone of desired effects" (the blue zone in the picture at right) – thus identifying those activities/interventions that were found to have large impacts on students' achievement outcomes (Hattie, 2009, p. 19). The infographic shown in Figure 1 visualizes the effects on student achievement from a variety of educational practices/activities/interventions.

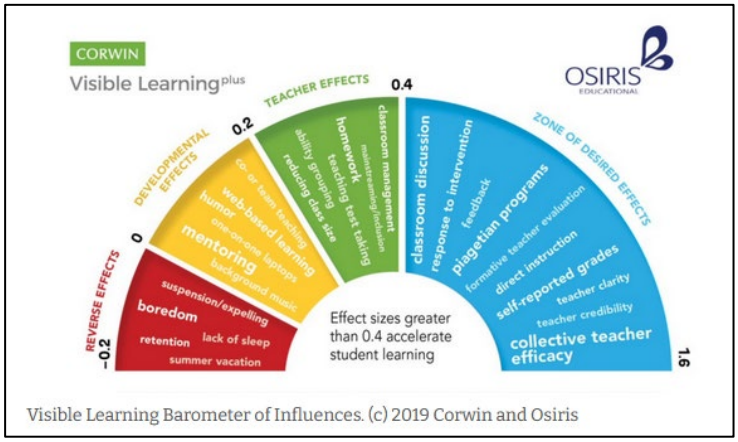

Figure 1. Hattie's Barometer[2] of Achievement Influences

Unlike the WWC, which defines and follows rigorous rules for identification and inclusion of studies in its meta-analyses, *Visible Learning* synthesizes results from existing meta-analyses, which may vary in quality, outcomes, comparison group programs/activities, and study inclusion criteria for each of the interventions reviewed. Thus, educators must be mindful of study features when interpreting the meaningfulness of these summary effect sizes. One recommendation is to follow-up information from *Visible Learning* with additional details from the WWC or other clearinghouses.

What Works Clearinghouse Improvement Index

To aid in summarizing effect sizes and comparing across different interventions – particularly for level 1 and 2 interventions that have been rigorously studied – the WWC uses an Improvement Index to convey an intervention's impact. The WWC Improvement Index translates effect sizes into a convenient and standard format. It summarizes the difference in percentile ranks between the average score of the intervention group and the average (50th percentile) score of the comparison group, according to the comparison group distribution. In this way, the Improvement Index estimates the amount of change in terms of percentile rank (on the noted outcome) that the average comparison group student would have experienced *if that student had received the intervention*.

[1] Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
[2] Infographic from: https://visible-learning.org/2022/01/hatties-barometer-of-influence-infographic/

For example, in Figure 2 below, a student in the comparison group who received the average score within the comparison group would be at the 50th percentile relative to students from this same group.

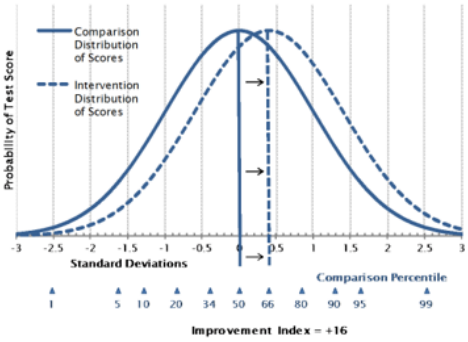

Figure 2: WWC Improvement Index Demonstration
Note: from WWC Procedures Handbook V4.1 (p.18)

Where does the average student from the intervention group fall based on the comparison group distribution? In this example, imagine the average score for the intervention group is .40 standard deviations above the average score for the comparison group, which also corresponds to the 66th percentile on the comparison group distribution. The Improvement Index is 16 points (66th percentile – 50th percentile). The value of 16 represents the expected change in percentile rank for an average comparison group student *if* that student had received the intervention. Using the expected change in percentile rank allows for a more concrete interpretation and comparable value of effect size.

In Figure 3, we provide two examples of how the Improvement Index is reported in WWC Intervention Guides for two different adolescent literacy interventions: *Achieve3000*[3] and *Peer-Assisted Learning Strategies*[4] (*PALS*). We see that *Achieve3000* had an average improvement index of +6 points for the domain of Comprehension, and +3 points for the domain of General Literacy Achievement. For *PALS*, the Comprehension domain has an average of +19 points. While these interventions targeted a different set of student outcomes, both targeted Comprehension, with *PALS* having a much higher average index. But note the number of studies and the number of students included in the respective intervention reviews. Since there was only one study of *PALS* deemed rigorous enough to be included in the WWC review, there is less information from multiple studies, which may affect educator confidence in these statistics. Overall, educators need to reflect on the entire corpus of information available in order to make a responsible intervention adoption decision.

Achieve3000

		Improvement index *(percentile points)*				
Outcome domain	**Rating of effectiveness**	**Average**	**Range**	**Number of studies**	**Number of students**	**Extent of evidence**
Comprehension	Potentially positive effects	+6	0 to +11	2	12,698	Medium to large
General literacy achievement	Potentially positive effects	+3	+2 to +3	2	32,110	Medium to large

Peer-Assisted Learning Strategies

		Improvement index *(percentile points)*				
Outcome domain	**Rating of effectiveness**	**Average**	**Range**	**Number of studies**	**Number of students**	**Extent of evidence**
Comprehension	Potentially positive effects	+19	na	1	120	Small

na = not applicable

Figure 3. Comparison of Adolescent Literacy Improvement Indices for *Achieve3000* and *PALS*.
Notes: Top: *Achieve3000* Intervention Guide (2018); Bottom: *Peer-Assisted Learning Strategies* (2012); refs. in footnotes

[3] Intervention Guide for Achieve3000 (2018): https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_alachieve_022718.pdf
[4] Intervention Guide for Peer-Assisted Learning Strategies (2012): https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_pals_013112.pdf

Summary
Effect sizes provide one piece of information regarding the promise of an intervention to influence achievement outcomes. There are also many different kinds of effect sizes, only a few of which were discussed here. Educators are urged to seek out as much information as possible for a given intervention – as well as the comparison it was tested against – in order to make the most effective choices for their students, classrooms and schools. To learn more about specific effect sizes or other issues related to their use, see our resources and references below.

Resources

REL-WEST quick-reference guide
https://ies.ed.gov/ncee/edlabs/regions/west/relwestFiles/pdf/4-2-3-14_Effect_Size_Infographic_Final_508c.pdf (2021) This quick-reference guide from the Institute of Education Sciences Regional Education Lab (REL-West) at Wested, provides a glossary of terms, dives a bit deeper into connections between effect size and statistical significance, and reviews additional literature on study and design features that can affect the size of an effect size statistic.

Scribbr
https://www.scribbr.com/statistics/effect-size/ Scribbr provides a simple description to effect sizes, similar to this guide, with additional links and examples.

Psychometrica
https://www.psychometrica.de/effect_size.html Psychometrika provides a web-based tool for calculation of many different kinds of effect sizes. It's designed for researchers who are conducting studies to compare outcomes across an intervention and comparison group.

Dr. Jerry Bean's effect size guide
A Guide to Common Effect Sizes and Forest Plots (2021) – Dr. Jerry Bean has created a handout describing the links between meta-analyses and effect sizes (used by permission).

WWC Procedures Handbook V4.1
https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC_Procedures_Handbook_V4_1_Draft.pdf
WWC Procedures Handbook V4.1 (p.18). This reference is used by trained specialists at the WWC who are tasked with conducting meta-analyses and determining effect sizes and improvement indices for catalogued interventions.

WWC Intervention Reports
https://ies.ed.gov/ncee/wwc/Search/Products?productType=2 Searchable catalog of interventions reviewed by WWC and for which sufficient and reliable information is available to create an Intervention Report.