Article

NASSP Bulletin
2019, Vol. 103(4) 273–280

DOI: 10.1177/0192636519889151
journals.sagepub.com/home/bul
SAGE

Interpreting Average Effect Sizes: Never a Center Without a Spread

Thomas R. Guskey[1,2]

Abstract

School leaders today are making important decisions regarding education innovations based on published average effect sizes, even though few understand exactly how effect sizes are calculated or what they mean. This article explains how average effect sizes are determined in meta-analyses and the importance of including measures of variability with any average effect size. By considering the variation in effect sizes among studies of the same innovation, education leaders can make better decisions about innovations and greatly increase the likelihood of achieving optimal results from implementation.

Keywords

effect size, meta-analysis, innovation, feedback

> There are no simple answers to complex problems.
>
> —Valerio Massimo Manfredi

Nearly every discussion about educational improvement today refers to "effect sizes." Education organizations compare effect sizes in planning professional learning programs. District and school leaders consider effect sizes when selecting the strategies to include in school improvement initiatives. Even classroom teachers evaluate effect sizes in deciding what practices will be most effective in helping their students learn.

What is odd about our fixation on effect sizes is that few educators know exactly what they are or how they are determined. Most practitioners have a general

[1]University of Louisville, Louisville, KY, USA
[2]University of Kentucky, Lexington, KY, USA

Corresponding Author:
Thomas R. Guskey, 2108 Shelton Road, Lexington, KY 40515, USA.
Email: guskey@uky.edu

understanding that effect size is a measure of "treatment effect." In education, this typically refers to an innovation's effectiveness in improving student learning. An innovation with an average effect size of +0.8, for example, is generally considered to be twice as effective as another innovation with an average effect size of only +0.4. Implementing the first innovation, therefore, will likely yield twice as much improvement in student learning as implementing the second. But is that really true?

Let us be clear: Effect size is a powerful tool when considering the value and effectiveness of various policies, strategies, practices, or innovations in education. But to use them appropriately in making major decisions about improvements in education, it is essential to know how they are determined, how to interpret them, and precisely what they mean.

What Is Effect Size?

Effect size is a statistic first described by psychologist Jacob Cohen in his book *Statistical Power Analysis for the Behavioral Sciences* (Cohen, 1969) and referred to by researchers as "Cohen's *d*." Originally developed to add substantive meaning to statistical significance testing, effect size was adapted for use in synthesizing results from multiple studies in "meta-analyses" (Bangert-Drowns, 1986). It provided researchers with a way to "standardize" the treatment effect in any study so that results from multiple studies conducted in different contexts and with different subjects can be compared or summarized. Effect sizes are especially prominent in the social sciences and in medical research where the magnitude of the treatment effect is particularly important.

Although there are several ways to determine the treatment effect in a scientific study, the most common is simply to compare the average score attained by subjects in the "treatment" group with the average score attained by those in a "control" group. In studying a particular teaching strategy, for example, we might compare the average score on a measure of achievement attained by students taught by a new strategy (i.e., treatment) with that attained by students who were taught by traditional methods or by an alternative strategy (i.e., control).

Suppose we made such a comparison and found that the average score of students who were taught by the new strategy was 10 points higher. That sounds terrific, of course. But how do we know if that difference is substantial or relatively modest? And how could we ever compare the 10-point difference in this study to another study of the same strategy conducted with different students and using a different measure of achievement?

Standardization

In order to bring meaning to this difference and make comparisons across studies, we need to convert this 10-point difference to a common, "standardized" metric that could be used for all studies. The procedure originally recommended by Cohen and later

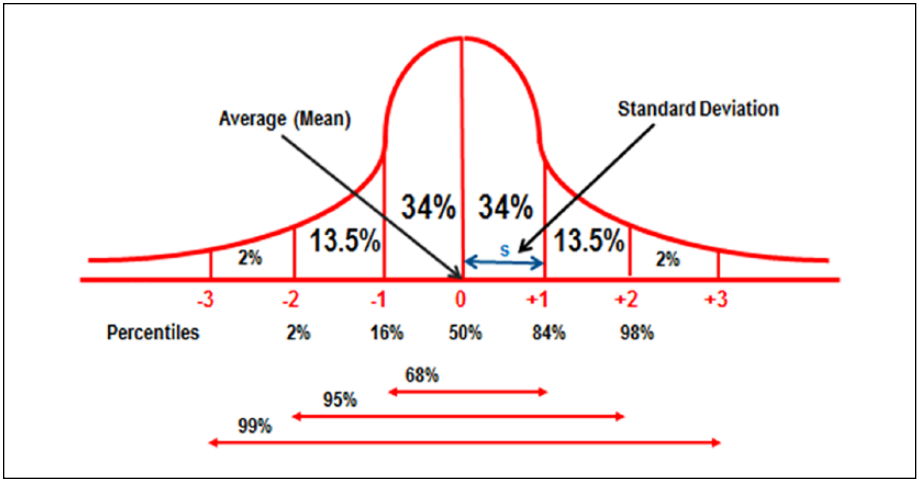

Figure 1. The normal curve distribution.

refined by Glass (1976) is based on a measure of the variation among scores in the control group, which is an estimate of the variation in scores in the population. Scores in the control group frequently resemble a normal distribution pattern like that shown in Figure 1. Most scores in the group clustered close to the average or "mean," with fewer scores occurring either far above or far below the average. A measure of the "typical" amount by which scores vary from the average is called the "standard deviation."

In normal distributions, approximately 68% of scores fall within 1 standard deviation above and below the average. About 95% of scores fall within 2 standard deviations above and below the average. Because standard deviations can be computed based on the variation in scores on any measure, they are used to "standardize" the size of the effect. In essence, effect size is a measure of the difference between groups in standard deviation units based on the variation of scores in the control group.

So let us go back to our example. Suppose the average score of students who were taught by the traditional strategy (i.e., control) was 60, and the average score of students taught by the new strategy (i.e., treatment) was 70, yielding the 10-point difference we described earlier. And suppose the standard deviation of the scores of students taught by the traditional strategy (i.e., control) was 10 points. This would mean that approximately 68% of traditional strategy students scored between 50 and 70; approximately 95% scored between 40 and 80. It also would mean that the 10-point difference achieved by new strategy students represents a 1 standard deviation positive difference. Hence, the effect size of the new strategy would be +1.0. Students who experienced the new strategy scored 1 standard deviation higher than students taught by the traditional strategy.

A 1 standard deviation improvement or effect size of +1.0 may seem modest. But in terms of a treatment effect in education, it is *huge*. Looking at Figure 2, we can see more precisely what a treatment effect of this size implies. In terms of percentiles, it

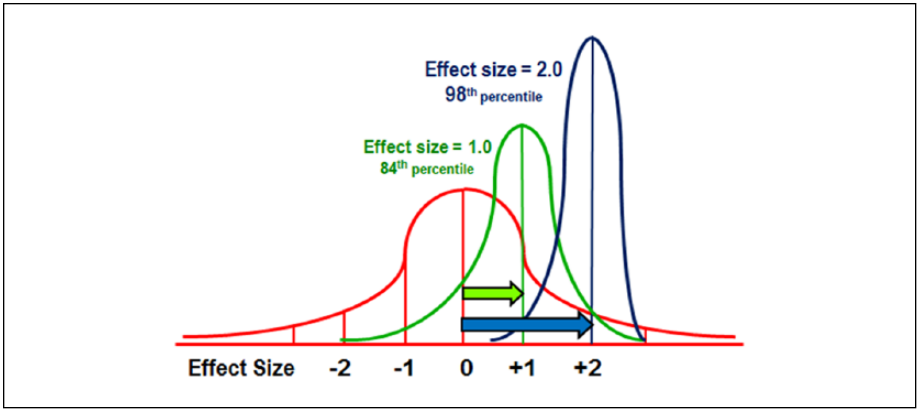

Figure 2. What effect sizes look like.

means that the average student in classes taught by the new strategy scored at a level achieved by only the top 16% of students in traditionally taught classes. An effect size of +2.0 would mean that the average student in the treatment class achieved at a level attained by only the top 2% of students in the control class. We do not know how long it took to achieve these results, but in this case, an effect size of +1.0 shows that the new strategy yielded an average 34 percentile improvement in student achievement.

Never a Center Without a Spread

Nearly everyone who pursues an advanced degree in education is required to take an introductory statistics course. Statistics, after all, is the language of research. It is how we summarize, analyze, and make sense of data in order to build knowledge and further our understanding.

One of the first topics discussed in introductory statistics courses is measures of central tendency and variability. We use these two measures to summarize any group of scores. A measure of central tendency is the score that typifies an entire group of scores. The most common measure of central tendency, of course, is the average or "mean." It is the score, usually in the middle of the group, around which the other scores "center."

Statistics teachers are quick to add, however, that you *never report a center without a spread*. In other words, to accurately describe any group of scores, the typical score or center *must* be accompanied by an indication of how much the other scores in the group vary from that center. This is called a measure of variability. As we described above, the most common measure of variability is the "standard deviation," which is generally interpreted as the "typical" amount by which other scores in the group differ from the average score or mean.

Including a measure of variability or spread is important because it provides an indication of how representative the average truly is. If the spread is small, then we

know the average is a fairly accurate representation of the group of scores. In other words, most of the scores in the group are fairly close to the mean. But if the spread is large, then we know the scores vary widely from that average score.

Meta-Analyses

Researchers who want to synthesize results from multiple studies of a particular treatment or innovation conduct "meta-analyses" in which they tally effect sizes from a collection of investigations conducted on the same policy, strategy, or practices but in different contexts with different individuals. In other words, they calculate an average or mean effect size by combining the individual effect sizes computed in each of the investigations assembled.

To meaningfully interpret this average, however, we also need a spread. We need to know if these studies all yielded similar effects sizes or if the effect sizes vary across studies. How much variation is there? And if the treatment or innovation was the same in each study—an assumption we make when we combine results in meta-analyses—then what could explain this variation?

The importance of considering variation in effect sizes is illustrated in a large-scale meta-analysis conducted by Kingston and Nash (2011) on effects of feedback provided through formative assessments in Grades K-12. They reviewed over 300 studies in their analysis but found that most had severely flawed research designs that yielded uninterpretable results. Only 13 studies provided sufficient information to calculate 42 independent effect sizes. The distribution of those effect sizes is shown in the stem-and-leaf plot in Figure 3. In this stem-and-leaf plot, the stem represents the units place and the tenths place of the effect size from each study and the leaf represents the hundredths place. So the first effect size reported at the top of the plot is for a study that yielded an effect size of −1.05; the fourth entry from the top represents three studies with effect sizes of −0.20, −0.24, and −0.26.

The Kingston and Nash (2011) meta-analysis yielded a median effect size of only +0.25, which challenged the results of earlier meta-analyses that estimated the average effect size for feedback from formative assessments to be between +0.70 and +0.90 (see Black & Wiliam, 1998; Hattie, 2009). Hence, instead of resulting in 30 to 40 percentile points average improvement, Kingston and Nash (2011) suggested the average improvement was only about 10 percentile points.

More important, however, Kingston and Nash (2011) considered the variation in effect sizes from study to study and found it was enormous. The 42 independent effect sizes ranged from −1.0 to +1.5. In other words, depending on the study, the impact varied from a decline of 35 percentile points to an increase of 43 percentile points!

Explaining Variation

Remember, the assumption in meta-analyses is that effect of the treatment—in this case, feedback from formative assessments—is consistent across studies. We simply could not make sense of any meta-analysis if the treatment being studied differed from

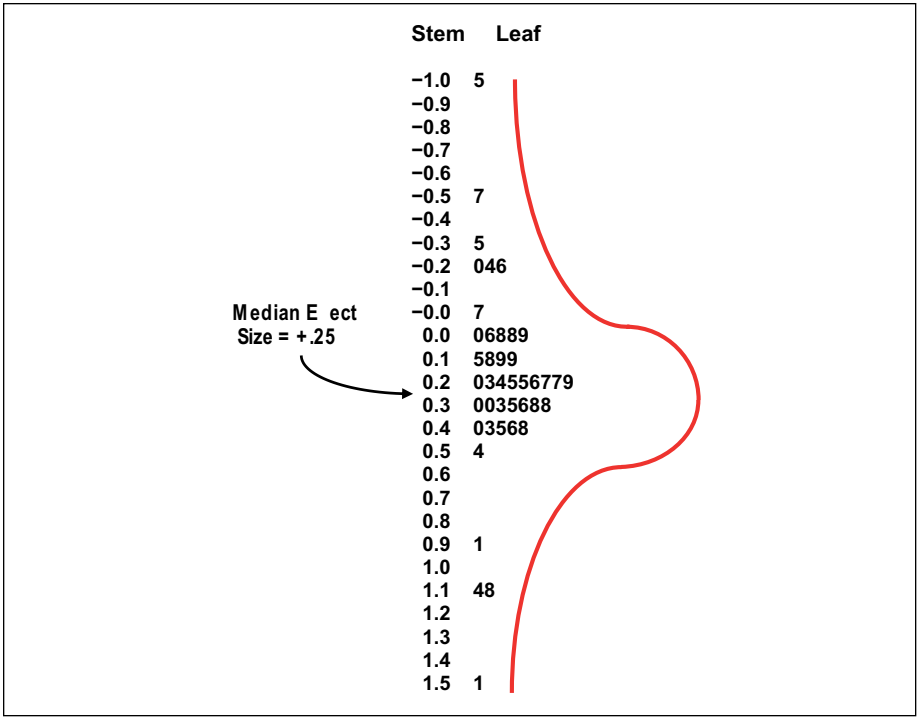

Figure 3. Stem-and-leaf plot of effect sizes from Kingston and Nash (2011).

study to study. Therefore, if the treatment is the same in each of these studies, then something other than the treatment must account for the tremendous variation in effect sizes. That is what Kingston and Nash (2011) set out to determine next. Based on information included in each study, they tried to ascertain what factors might explain the large variation in effect sizes.

They discovered that a small portion of the variation (about 2%) was attributable to differences in grade level. The effect of feedback from formative assessments was slightly more powerful in lower grade levels than in upper grades. The way formative assessments were implemented accounted somewhat more variation (about 15%), with professional development for teachers and the use of computer-based formative systems being more effective than other approaches.

The largest portion of the variation (about 58%) was due to subject area differences. Feedback from formative assessments was generally more effective in English language arts than in mathematics or science, with estimated group effect sizes of +0.32, +0.17, and +0.19, respectively (Kingston & Nash, 2015). So the effects of formative assessment feedback appear to differ depending on the grade level of students, the way it is implemented, and especially the subject area of instruction. Their conclusion about the true impact of feedback from formative assessments on student learning was essentially, "It depends."

Although some may consider the Kingston and Nash (2011) analysis an anomaly, other reviews of research on the effects of feedback in general verify how complex the effects can be. Fyfe and Rittle-Johnson (2016), for example, found that feedback can both help and hinder learning. Their analysis revealed that despite broad endorsement of feedback, research indicates the effects of feedback vary considerably depending on students' prior knowledge and are not universally beneficial (see also Mory, 2004).

Implications

So what does this mean for busy education practitioners who are looking for guidance in selecting policies, strategies, practices, or innovations that will best help them improve student learning? First, they must recognize that *average effect sizes alone are not enough*. Accuracy in educational measurement and correctness in interpreting the results of educational research demand measures of variability. *Never a center without a spread!* To accurately interpret the average effect size from any meta-analysis, or analysis of meta-analyses, requires an accompanying measure of variability. We cannot judge the true meaning of that average without it. There are no exceptions.

Second, if the variability of effect sizes in any meta-analysis, or analysis of meta-analyses, is significant, then efforts should be made to explain that variability. If effects are inconsistent across studies, practitioners need to know what factors explain that variation. They need to know, for example, if effects vary depending on student characteristics such as age or grade level, gender, or academic or cultural background. They need to know if they should expect different results depending on characteristics of the teachers involved, the subject area, the school, or the community. Too often, educators are led to believe that if they implement a particular innovation and do not see the same magnitude of effect size as described in popular publications, they must be doing something wrong. But that may not be the case.

Finally, practitioners must build in procedures to gather evidence of the effects on students of any policy, strategy, set of practices, or innovation they plan to implement. This should be evidence that teachers trust and that will help teachers determine if they are achieving the magnitude of improvement they hoped to see and were led to expect. More important, such evidence will help them identify problems and difficulties that may need to be addressed in order to achieve the results they want.

Average effect size is a vital statistic that helps educators make sense of syntheses of research on different educational policies, strategies, practices, and innovations. But when used to describe the effects of any treatment, it tells only half the story. The other half comes from a measure of the variability that shows how much the effect sizes of individual studies fluctuate around that average. *Never a center without a spread!* Both statistics are necessary to adequately describe meta-analytic results. If the variation in effect sizes proves to be significant, it means that factors other than the treatment are influencing the results, and additional steps must be taken to explore precisely what those factors might be. A measure of variability is crucial in interpreting meta-analyses results and essential to understanding what that average effect size really means.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Thomas R. Guskey https://orcid.org/0000-0003-1383-2407

References

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin*, *99*, 388-399.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139-144.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Lawrence Erlbaum.

Fyfe, E. R., & Rittle-Johnson, B. (2016). Feedback both helps and hinders learning: The causal role of prior knowledge. *Journal of Educational Psychology*, *108*, 82-97.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28-37.

Kingston, N., & Nash, B. (2015). Erratum. *Educational Measurement: Issues and Practice*, *34*(2), 55.

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology: A project for the Association for Educational Communications and Technology* (2nd ed., pp. 745-783). Mahwah, NJ: Lawrence Erlbaum.

Author Biography

Thomas R. Guskey is senior research scholar at the University of Louisville and Professor Emeritus at the University of Kentucky. A graduate of the University of Chicago, his research focuses on professional learning and teacher change; evaluation; and assessment, grading, and reporting student learning. Contact him at guskey@uky.edu or www.tguskey.com.