Beyond small, medium, or large: Points of consideration when interpreting effect sizes

Arthur Bakker, Jinfa Cai, Lyn English, Gabriele Kaiser, Vilma Mesa, & Wim Van Dooren

Suppose that researchers are interested in determining whether there is a group difference in an outcome as a result from an intervention. At least three questions are then worth asking:

> (a) Is an observed effect real or should it be attributed to chance?
>
> (b) If the effect is real, how large is it?
>
> (c) Is the effect large enough to be useful? (Kirk, 2001, p. 213)

The first question has been the topic of much debate, particularly in relation to the use and misuse of *p* values (Nuzzo, 2014; Trafimow & Marks, 2015; Wasserstein & Lazar, 2016). In this editorial we focus on effect sizes, descriptive statistics that can assist in answering the second and third questions.

Scientists have discussed the use of effect sizes to report treatment effect in various fields, such as medicine (Sullivan & Feinn, 2012), education (Baird & Pane, 2019; Lipsey et al., 2012; Shadish et al., 2015), and psychology (Prentice & Miller, 1992; Wilkinson & the APA Taskforce, 1999). While the APA Publication manual suggests that reporting effect sizes for primary findings is a must in addition to *p* values (APA, 2010), currently journals in the domain of mathematics education are not explicit about reporting or interpreting effect sizes. The purpose of this editorial is to draw researchers' attention to the importance of effect sizes in quantitative studies, as well as to stimulate sensible interpretation of them beyond the commonly used characterizations in terms of small, medium, or large. As Lipsey et al. (2012) wrote: "The widespread indiscriminate use of Cohen's generic small, medium, and large effect size values to characterize effect sizes in domains to which his normative values do not apply is [...] inappropriate and misleading" (p. 4). To avoid such uncritical use of tables with benchmark values, we highlight twelve points of consideration when interpreting the magnitude and relevance of a difference or association. We also make a call for alternative guidelines that can be used in our discipline of mathematics education research.

1. What is an effect size?

An effect size provides a quantitative measure of the magnitude of the difference between groups or association between variables. It provides an assessment of the strength of findings that tests of statistical significance alone do not provide (Balow, 2017; Coe, 2002; Durlak, 2009; Sullivan & Feinn, 2012). There are many different measures of effect size appropriate for different situations (Kirk, 1996, counted 40 measures). They belong to one of two families: difference or correlation (association) measures (Ellis, 2010). The most common standardized difference measures are that of a proportion of a standard deviation (SD), for example "0.21 SD," and Cohen's $d = |M_1 - M_2|/SD$ (where the SD is usually the pooled SD).

Division by the SD makes the measure standardized (independent of a unit), which allows comparison across studies. Cohen's *d* is used in many meta-analyses, and in syntheses of meta-analyses (e.g., Hattie, 2009). Correlational effect sizes such as r^2 and Eta^2 are interpreted somewhat differently: They refer to the percentage of explained variance.

2. Why report effect sizes in quantitative studies?

It has become preferred practice to report effect sizes in quantitative studies in the social sciences (Durlak, 2009; Coe, 2002; Sun, Pan, & Wang, 2010). Wilkinson et al. (1999) go so far to advise: "Always report effect sizes for primary outcomes" (p. 599). While tests for statistical significance simply tell the reader that the difference found between groups is unlikely to be caused purely by chance, the *p* value does not tell how large the difference is (Balow, 2017; Durlak, 2009; Sullivan & Feinn, 2012; Sun et al., 2010). Some editors have even banned *p* values (null hypothesis significance testing) from their journal (Trafimow & Marks, 2015). If the *p* value is below .05 but the effect size is negligible, it may not be worth investing in the intervention or drawing conclusions for theory development. It is important to keep sample size into account when deciding on practical or theoretical relevance. For example, with very large sample sizes, it is likely that a statistically significant difference will be found. When working with big data sets, *p* values are therefore less informative, and effect sizes are more relevant because the latter can help decide whether the difference found is meaningful or not.

Reporting effect sizes is also important because the values reported help other researchers conduct power analyses to determine the appropriate sample size (Cohen, 1969; Sun et al., 2010). Moreover, reporting effect sizes, in addition to means and standard deviations, allows other researchers to conduct meta-analyses (Ellis, 2010).

3. The need for updating guidelines for interpreting effect sizes

Fifty years ago, Cohen (1969) developed benchmark values for the effect size *d* (which he called an index), in the context of small-scale experiments in social psychology. The benchmark values are widely used today: 0.2 small, 0.5 medium, and 0.8 large. While Cohen set the values because there were no guidelines then, he warned that they were somewhat arbitrary. Moreover, he stressed that the benchmark values were "recommended for use only when no better basis for estimating the index is available" (1988, p. 25).

Those benchmark values seem to have had merit in educational research too: In a meta-analysis of 300 experiments, Lipsey and Wilson (1993) found an average effect size of 0.5 SD (half a standard deviation unit), suggesting this value was indeed medium. But effect sizes have decreased over time. Lipsey et al. (2012) analyzed 124 randomized controlled trials (RCTs) and found a much lower average effect size of 0.28 SD. More recently, even smaller averages have been found: 0.16 SD on the basis of 197 RCTs on achievement (Cheung & Slavin, 2016); 0.05 SD for mathematics and 0.07 SD for reading on the basis of 105 school-wide RCTs (Fryer, 2017). Lortie-Forgues and Inglis (2019) found an average of 0.06 SD for 141 RCTs in the US and UK, with large confidence intervals. It is up for discussion why these averages seem to have gone down, and whether such averages are informative.

Irrespective of how they are explained, however, the decreasing average effect sizes suggest that the benchmark values need updating. Over the past five decades, it has become clear that there are many points that need to be considered if the size and significance of effect sizes are to be interpreted. There is thus a need for alternative and more refined interpretations of effect sizes.

4. Twelve points to consider when interpreting effect sizes

The points to be taken into account are numerous (e.g., Cheung & Slavin, 2016; Kraft, 2019). We suggest twelve points for authors and reviewers in mathematics education journals to consider when interpreting effect sizes. Some points are technical, depending on the formula used (1–3). Others are primarily methodological (4–8). The remaining points seem to be empirical or perhaps even ontological. Explanations of why these points influence effect sizes are not always so clear, in particular when it comes to ontological ones, hence further discussion is welcome.

1. *Which calculation of the effect size is used?* Different computations are available, and there is discussion about which SD has to be used. The values produced inevitably depend on choices made on how to compute an effect size (Kraft, 2019).
2. *Is there very small variance?* The formulas for the effect sizes in the difference family include a measure of variance. For example, the formula of Cohen's *d* implies that very small variance, hence a very small SD, can have a huge influence on the value of *d,* because a difference is divided by the SD. If for instance student scores on a pretest are all close to the bottom value, the value of *d* will be artificially high.
3. *What is the confidence interval around the point estimate?* An effect size is a point estimate, often with rather large confidence intervals (Lortie-Forgues & Inglis, 2019; Zieffler, Harring, & Long, 2011).
4. *To what is the effect compared?* Effect size is a relational measure, therefore it matters what is compared with what: The change in the experimental condition versus the change in the control condition? One experimental condition with another one? Within-group change from pre- to posttest? The latter difference can be expected to be larger than the change in the experimental condition minus the change in the control condition, hence it matters what is compared to what.
5. *What is the research design?* Effect sizes found in RCTs tend to be much smaller than in quasi-experimental designs (Cheung & Slavin, 2016). It is up for discussion whether effect sizes should also reported for differences between pre- and posttest results. These within-group changes (pre-post within one group) tend to be larger than the between-group differences in the changes in two conditions. Some scholars discourage reporting effect sizes for pre-posttest designs, for example because pre- and posttest results tend be dependent (Cuijpers et al., 2017; Kraft, 2019). Originally, effect sizes were used for experimental studies, so that they were measures of causal effects (compared with a control group).
   The use of effect sizes has been expanded to correlational and associational measures (Ellis, 2010). A possible advantage of correlational effect sizes such as r^2 and Eta^2 is

that they refer to the percentage of explained variance. “Effect sizes” from correlational studies tend to be larger than those from experimental studies (Lipsey & Wilson, 1993; Lipsey et al., 2012), yet should not be interpreted as causal effects because the study design is correlational in nature. These observations imply that the magnitude of an effect size should be considered in relation to the research design.

6. *To what extent are the intervention and measurement aligned?* The more the test resembles what is taught, the larger the effect size one can expect (Cook et al., 2015). In an analysis of 645 studies, Cheung and Slavin (2016) found that effect sizes were about twice as large in studies with researcher-developed (“treatment-inherent” or “experimenter-made”) measures than if independent measures were used. This is not to problematize researcher-made tests as their validity depends on whether they measure what they purport to measure. Yet it is relevant to know what kind of test was used in order to judge the magnitude of an effect size.
7. *Focus on offering or receiving?* Even if an intervention is offered to many, only some may actually take it up. In an intensive tutoring study, a 0.23 SD effect size was found on mathematics achievement for those who actually received the tutoring, whereas the effect size was only 0.13 SD when computed for those to whom the tutoring was offered (Kraft, 2019).
8. *How intensive or long was the intervention?* Some results may be the outcome of short-term decisions whereas others are the cumulative effect of sustained effort (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). If an intervention is longer or more intense, with more opportunities to learn, one can expect larger effect sizes because of cumulative effects. However, there may be a limit to this because in the long run, it may be harder to control conditions. It is therefore hard to judge on the magnitude of an effect size if no information is known about length or intensity of an intervention.
9. *What is the sample size?* Cheung and Slavin (2016) found that average effect size in studies with sample size up to 100 was about 3.5 times larger than in studies with large samples (2000+): 0.38 vs 0.11. Hence, sample size also matters in interpreting effect sizes.
10. *What was the sample?* Interventions on particular groups such as historically marginalized students tend to lead to larger effect sizes than interventions with, say, privileged groups who already score rather high (Paunesku et al., 2015). Relatedly, heterogeneous groups (in terms of the variables at interest) may be harder to influence than homogeneous ones (Kraft, 2019). However, larger effects found in homogeneous groups may also be an artifact of having smaller variation (see the first point in this list). It has further been observed that younger children may be easier to influence than older ones (Bloom et al., 2008), for example when it comes to achievement and attitudes toward mathematics and science (Savelsbergh et al., 2016).
11. *How easily can one influence the dependent variable?* Some things are easier to influence than others. For example, in their meta-analysis on teaching approaches in science and mathematics education, Savelsbergh et al. (2016) found larger effects on achievement than on affect variables such as attitudes, interest, and motivation.

12. *What is the context?* If education in a particular context is already of a high standard, it may be more difficult to achieve large effect sizes (ceiling effect) than in educational systems with more room for improvement. Moreover, it may be easier to have some control in national curricula or areas with centralized policy structures than in decentralized educational systems where teachers have a lot of agency.

When it comes to practical relevance, it is even more challenging to interpret effect sizes. Based on effect sizes, will people be able to judge in what program they could best invest? Isn't this too much for one number to carry? Hattie (2009) aimed to provide an overview of what works to promote achievement, based on effect sizes reported in meta-analyses. In his book, he acknowledges some of the aforementioned points of consideration (Snook et al., 2009), but his lists of effect sizes ignore these points and are therefore misleading (for elaborate critiques of his way of using effect sizes see Bergeron & Rivard, 2017; Terhart, 2011). When it comes to practical significance, two of the key factors that policymakers need to know in addition to effect size are scalability and costs (Harris, 2009). Some interventions are scalable and cheap, but many are expensive. In addition, some interventions with smaller effects can be easily combined, while others may contradict each other (Kraft, 2019).

5. A call for better guidelines

From the list of points, it will be clear that it does not make sense to judge the magnitude of an effect size on the basis of commonly used tables with benchmark values including Cohen's original ones. Let us take a first virtual example of an RCT, with a large sample (2000+), short intervention, cheap, dependent affect variable (e.g., motivation, interest, attitudes), or a standardized test (not experimenter-made). On the basis of Cheung and Slavin's (2016) overview, Cohen's $d = 0.15$ (0.15 SD) might be impressive, even though in commonly used benchmark tables it would be not even considered a small effect. Let us contrast this with another virtual example: Pre-post measurement of a specific skill using a researcher-developed test, in an intensive intervention dedicated to this skill, and a small sample. Though ten times larger numerically, Cohen's $d = 1.5$ (1.5 SD) may be disappointing, even though it is much better than what is often characterized as a large effect (0.8+). In some people's view such effect size should not even be reported given that it is not a comparison with another condition (Kraft, 2019). Kraft (2019) proposes a new approach for causal studies on achievement, a matrix including cost and scalability. Although his proposal is an improvement for policy decisions, we think that an even more situated judgment of effect sizes is needed when reporting them in the field of mathematics education.

One important aspect of situating the research is realizing that the role and interpretation of effect size will depend on the goal of the study. Most of the comments above relate to studies that investigate the effectiveness of a specific intervention, with the idea that this intervention—when effective—can be implemented in mathematics education practice later on. But particularly in education, there are also other kinds of (quantitative) empirical studies where this is not necessarily the aim (Mook, 1983). In these studies, the claim under investigation is not a claim about the size of the effect of a specific intervention or a specific variable. Rather, a theoretical claim is tested that X influences Y. The idea behind such

studies is that experimental research can be used to provide evidence for a theoretical claim, and thus enhance our understanding of mathematical thinking, learning, and teaching. A multitude of such experiments may then lead to a theory substantiated by empirical evidence. It may be argued that in such cases the size of an effect is less important than when the intention is to implement an idea at scale.

The bottom line of our argument is that when effect sizes are reported, they need a situated interpretation with awareness of the aforementioned points of consideration. Some authors report effect sizes without interpreting them as small, medium, or large. Although this is, in our view, better than uncritically calling on a benchmark values such as the aforementioned 0.2 – 0.5 – 0.8, we encourage authors first of all to judge if reporting effect sizes make sense in their situation, and to interpret the obtained effect size in light of the goals of their study. Our advice is that these effect sizes are compared to effect sizes reported in similar studies in terms of "smaller/larger than typical under such conditions," or "comparable with other studies with similar characteristics (research design, alignment between intervention and assessment, sample size, type of variable influenced etc.)." It may not always be easy to find comparable studies, yet we think such comparison is necessary.

To summarize our messages:

1. Report effect sizes when presenting the primary outcomes of (quasi-)experimental studies, in line with APA (2010) guidelines. We do not propose a one-size-fits-all approach where effect sizes should always be reported.
2. Do not characterize effect sizes on the basis of standard tables that call them small, medium, or large irrespective of the aforementioned points of consideration. Instead, relate your findings to comparable studies with similar characteristics (research design, sample size, type of measurement, type of variable influenced etc.).
3. We welcome overviews and careful interpretations of effect sizes found in mathematics education that include aforementioned points of consideration (and possibly others), so that other researchers can judge better how effect sizes found in other studies compare to their own. Such overviews can form the basis for new guidelines to be used in our discipline of mathematics education research.

Note on the authors

Arthur Bakker is Editor-in-Chief of *Educational Studies in Mathematics*; Jinfa Cai is Editor-in-Chief of *Journal for Research in Mathematics Education*; Lyn English is Editor-in-Chief of *Mathematical Thinking and Learning*; Gabriele Kaiser is Editor-in-Chief of *ZDM Mathematics Education*; Vilma Mesa and Wim Van Dooren are associate editors of *Educational Studies in Mathematics.*

Acknowledgments

We thank the following colleagues for their helpful suggestions: Rolf Biehler, Robert DelMas, Paul Drijvers, Rob Gould, William Penuel, Stanislaw Schukajlow, Sietske Tacoma, Chris Wild, and Andy Zieffler; and Nathalie Kuijpers for her editorial support.

References

APA (2010). *Publication manual of the American Psychological Association (*6th Ed.). Washington, DC: APA.

Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher*, *48*(4), 217–228. doi: 10.3102/0013189X19848729

Balow, C. (2017). The "effect size" in educational research: What is it and how to use it? *Illuminate Education*. Retrieved from www.illuminateed.com/blog/2017/06/effect-size-educational-research-use/ on July 14, 2019.

Bergeron, P., & Rivard, L. (2017). How to engage in pseudoscience with real data: A criticism of John Hattie's arguments in visible learning from the perspective of a statistician. *McGill Journal of Education / Revue des sciences de l'éducation de McGill*, *52*(1), 237–246. doi: 10.7202/1040816ar

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.

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292.

Coe, R. (2002). *It's the effect size, stupid. What effect size is and why it is important.* Paper presented at the British Educational Research Association annual conference, Exeter, UK. Retrieved July 24, 2019 from https://www.cem.org/attachments/ebe/ESguide.pdf

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1st Ed.). New York, NY: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Erlbaum.

Cook, P. J., Dodge, K., Farkas, G., Fryer, R. G., Guryan, J., Ludwig, J., & Mayer, S. (2015). *Not too late: Improving academic outcomes for disadvantaged youth.* Institute for Policy Research Northwestern University Working Paper WP-15-01.

Cuijpers, P., Weitz, E., Cristea, I. A., & Twisk, J. (2017). Pre-post effect sizes should be avoided in meta-analyses. *Epidemiology and Psychiatric Sciences*, *26*(4), 364–368.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, UK: Cambridge University Press.

Fryer Jr, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In E. Duflo & A. Banerjee (Eds.), *Handbook of economic field experiments* (Vol. 2, pp. 95–322). Amsterdam, the Netherlands: North-Holland.

Harris, D. N. (2009). Toward policy-relevant benchmarks for interpreting effect sizes: combining effects with costs. *Educational Evaluation and Policy Analysis*, *31*(1), 3–29. doi: 10.3102/0162373708327524

Hattie, J. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. Abingdon, UK: Routledge.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61*(2), 213–218.

Kraft, M. (2019). *Interpreting effect sizes of education interventions.* (EdWorkingPaper: 19-10). Retrieved from Annenberg Institute at Brown University: http://edworkingpapers.com/ai19-10

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*(12), 1181.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. Washington, DC: National Center for Special Education Research.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, *48*(3), 158–166.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*(4), 379–387. doi: 10.1037/0003-066X.38.4.379

Nuzzo, R. (2014), Scientific method: Statistical errors, *Nature*, 506, 150–152. http://www.nature.com/news/scientific-methodstatistical-errors-1.14700

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science*, *26*(6), 784–793.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*(1), 160.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *39*(5), 369–393.

Savelsbergh, E.R., Prins, G.T., Rietbergen, C., Fechner, S., Vaessen, B.E., Draijer, J.M. & Bakker, A. (2016). Effects of innovative science and mathematics teaching on student attitudes and achievement: A meta-analytic study. *Educational Research Review*, *19*, 158–172.

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The role of between-case effect size in conducting, interpreting, and summarizing single-case research* (NCER 2015-002). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Snook, I., O'Neill, J., Clark, J., O'Neill, A. M., & Openshaw, R. (2009). Invisible learnings?: a commentary on John Hattie's book-Visible learning: a synthesis of over 800 meta-analyses relating to achievement. *New Zealand Journal of Educational Studies*, *44*(1), 93–106.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. doi: 10.4300/JGME-D-12-00156.1

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, *102*(4), 989–1004. doi: 10.1037/a0019507

Terhart, E. (2011) Has John Hattie really found the holy grail of research on teaching? An extended review of Visible Learning. *Journal of Curriculum Studies*, *43*(3), 425–438. doi: 10.1080/00220272.2011.576774

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology 37*, 1–2.

Wasserstein, R. L., & Lazar, N. A. (2016) The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. doi: 10.1080/00031305.2016.1154108

Wilkinson, L., & APA Taskforce (1999). Statistical methods in psychology journals. *American Psychologist*, *54*(8), 594–604.

Zieffler, A. S., Harring, J. R., & Long, J. D. (2011). *Comparing groups: Randomization and bootstrap methods using R*. Hoboken, NJ: Wiley.