Behav Res (2013) 45:968–971
DOI 10.3758/s13428-013-0392-4

BRIEF COMMUNICATION

Cohen's *d* needs to be readily interpretable: Comment on Shieh (2013)

Geoff Cumming

Published online: 4 September 2013

Abstract Shieh (2013) discussed in detail δ^*, a proposed standardized effect size measure for the two-independent-groups design with heteroscedasticity. Shieh focused on inference—notably, the large challenge of calculating confidence intervals for δ^*. I contend, however, that the standardizer chosen for δ^*, meaning the units in which it is expressed, is appropriate for inference but causes δ^* to be inconsistent with conventional Cohen's d. In addition, δ^* depends on the relative sample sizes in the particular experiment and, thus, lacks the generality that is highly desirable if a standardized effect size is to be readily interpretable and also usable in meta-analysis. In the case of heteroscedasticity, I suggest that researchers should choose as standardizer for Cohen's δ the best available estimate of the *SD* of an appropriate population, usually the control population, in preference to δ^* as discussed by Shieh.

Keywords Cohen's d · Standardized mean difference · Effect sizes · Interpretation · Meta-analysis

Standardized effect sizes (ESs)—notably, Cohen's d—are widely used in psychology for two main reasons. First, a d value affords understanding and interpretation independently of the original measure and situation. A reduction in anxiety of $d = 0.3$ has at least some meaning, whatever the original anxiety scale and whatever the study in which it was measured. Second, most meta-analyses in the social and behavioral sciences need to use standardized ESs because the separate studies used a variety of original measures for the effect of interest—perhaps the reduction in anxiety produced by some psychotherapeutic procedure. Both of these two valuable features of d rely on the measurement unit of d—the standardizer—being chosen appropriately. Only with a consistent and readily interpretable standardizer can d have the generality—across researchers, experimental designs, and situations—that is necessary for meaningful interpretation and inclusion in meta-analyses.

Shieh (2013) recognized that "the choice of a suitable effect size parameter and statistic is a difficult and substantive decision" (p. 2). An important thread through the current comment is that, indeed, a range of substantive decisions are needed when selecting, calculating, and interpreting standardized ESs; informed judgment in context is required. Shieh's main purpose was to discuss a form of d proposed by Kulinskaya and Staudte (2007) for the case of two independent groups with heteroscedasticity. Shieh, like Kulinskaya and Staudte, focused on inference—notably, the challenging task of calculating confidence intervals (CIs). I contend, however, that the form of d studied by Shieh lacks generality, is inconsistent with familiar, widely used forms of d, and is likely not to be interpretable in any practically useful way.

Cohen's *d*

First, I briefly review some issues about d. I discuss these and other aspects of d in more detail in Cumming (2012, Chap. 11), which is accompanied by software (ESCI, *Exploratory Software for Confidence Intervals*, freely available from www.thenewstatistics.com) designed to support understanding of d and calculation of d and CIs on d in a range of common situations.

Cohen's d is a ratio: an ES divided by a standardizer, where both numerator and denominator are expressed in original units. Cohen's δ, the standardized mean difference between the population means, is given by Eq. 1, where μ_1 and μ_2 are

G. Cumming (✉)
Statistical Cognition Laboratory, School of Psychological Science, La Trobe University, Melbourne, Victoria 3086, Australia
e-mail: g.cumming@latrobe.edu.au

the two population means and σ, the standardizer, is the assumed common population *SD*:

$$\delta = (\mu_1 - \mu_2)/\sigma. \quad (1)$$

We use Eq. 2 to calculate d from our data, where d is our estimate of δ, M_1 and M_2 are the two group means, and s is our chosen standardizer:

$$d = (M_1 - M_2)/s. \quad (2)$$

As standardizer, s, we need to choose an estimate of σ; usually, s_p, the pooled within-groups *SD*, is best if we are happy to assume homogeneity of variance. However, despite the lower number of degrees of freedom and, thus, loss of precision, we might prefer s_1 if group 1 is a control group and we suspect heteroscedasticity because the experimental treatment may give a different, usually larger, *SD* in group 2.

Beyond those basics, we should recognize several challenging aspects of d. First, d calculated using Eq. 2 is the ratio of two quantities both subject to sampling variability. When considering any value of d, when comparing any two d values, or when conducting a meta-analysis using d, we must remember that all values are produced by an original-units ES value (the numerator in Eq. 2) and an estimated *SD* (the denominator). The standardizer (the denominator) is the measuring unit for d but will almost certainly be different for different samples from the same population, merely because of sampling variability. Cohen's d is thus measured on a *rubber ruler*, so called because the measuring scale stretches in or out each time we measure (Cumming, 2012, Chap. 11 explained further and provided ESCI simulations to illustrate). Great caution is needed when interpreting d.

Second, although my focus in this comment is on meaningfulness and interpretation, inference is always important—notably, to calculate CIs for d and to provide a variance estimate for use in meta-analysis. For the homogeneous case, using s_p as standardizer, an iterative procedure based on noncentral t distributions provides accurate CIs on d (Cumming, 2012, Chaps. 10–11; Cumming & Finch, 2001). Cumming and Fidler (2009) described and evaluated an excellent approximation. If we elect to use s_1, the *SD* of the control group, as standardizer, perhaps because we suspect heterogeneity, Hedges (1981, pp. 110–111) explained the calculations needed, again using noncentral t. Calculation of d in this case (as in the homogeneous case) is not sensitive to the relative sizes of the two samples, nor is the CI on d in the homogeneous case. However, it is interesting to note that the CI on d calculated using s_1 as standardizer—the heterogeneous case—*is* sensitive to the relative sizes of the two samples (Hedges, 1981, pp. 110–111). (I thank an anonymous reviewer for pointing this out.)

Third, d is a biased estimate of δ, being somewhat too large, especially for small samples. The correction needed to calculate d_{unb}, the unbiased version of d, was explained by Hedges (1981, pp. 112–116) and Cumming (2012, pp. 294–295). Unbiasing is easily accomplished as described in those sources or by using ESCI; d_{unb} should probably be our routine choice of standardized ES measure.

Note that I am using "Cohen's d" as a generic term for the standardized mean difference and d_{unb} for the unbiased version. Using Cohen's d accords with contemporary practice but requires that any mention of d must be accompanied by an explanation of how d was calculated, especially what standardizer was used. Various other terms, such as "Hedges's g," are also used, but with striking inconsistency (Cumming, 2012, pp. 295–296), and so are best avoided.

Choice of standardizer

Equations 1 and 2 show that δ and d are expressed in *SD* units: The standardizer is the unit of measurement. We need to choose as standardizer a population *SD*, σ, that makes conceptual sense (Kline, 2013, p. 133)—that can be given a substantive interpretation in context. Then we choose our best estimate of that σ to use as s in Eq. 2.

In the case of two independent groups, s_p is our likely choice of standardizer, and s_p is also used to conduct inference about the difference between the two group means. Whether applying an independent-groups t test or calculating a CI on the difference, we use s_p. In many cases, however, the best choice of standardizer is *not* the *SD* needed to conduct inference on the effect in question. Consider, for example, the paired design, such as a simple pre–post experiment in which a single group of participants provide both pretest and posttest data. The most appropriate standardizer is virtually always (Cumming, 2012, pp. 290–294; Cumming & Finch, 2001, pp. 568–570) an estimate of the *SD* in the pretest population, perhaps s_1, the pretest *SD* in our data. By contrast, inference about the difference requires s_{diff}, the *SD* of the paired differences—whether for a paired t test or to calculate a CI on the difference (Cumming & Finch, 2005). To the extent the pretest and posttest scores are correlated, s_{diff} will be smaller than s_1, our experiment will be more sensitive, and a value of d calculated erroneously using s_{diff} as standardizer will be too large.

The primary reason for choosing s_{pre} as standardizer in the paired design is that the pretest population *SD* virtually always makes the best conceptual sense as a reference unit. Another important reason is to get d values that are likely to be comparable to d values given by other paired-design experiments possibly having different pretest–posttest correlations and by experiments with different designs, including the independent-groups design, all of which examine the same effect. The d values in all such cases are likely to be comparable because they use the same standardizer—the control or

pretest *SD*. Such comparability is essential for meta-analysis, as well as for meaningful interpretation in context.

Shieh's δ^*

For the case of two independent groups of sizes N_1 and N_2, Shieh, following Kulinskaya and Staudte (2007), defined δ^*, given by Eq. 3, as the population standardized ES of interest, where $N = N_1 + N_2$, $q_1 = N_1 / N$, and $q_2 = N_2 / N$:

$$\delta^* = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/q_1 + \sigma_2^2/q_2}}. \quad (3)$$

Shieh provided a detailed discussion of δ^*, especially calculation of CIs on δ^* and the determination of sample sizes so an experiment is likely to give CIs on δ^* that are no longer than a stated length. He considered, in particular, cases having heteroscedasticity ($\sigma_1^2 \neq \sigma_2^2$) and unequal group sizes ($N_1 \neq N_2$). He provided valuable results and recommendations, especially in relation to inference based on δ^*. He made little mention, however, of how d^* values (estimates of δ^*) might be interpreted in practice. He stated clearly that δ^* is dependent on aspects of the experiment, notably q_1 and q_2, the relative sample sizes. The denominator in Eq. 3 is appropriate for Shieh's inferential purpose of calculating CIs but, I contend, has two serious deficiencies as a standardizer. Just as s_{diff} is what we need for inference in the paired design but should not be used as the standardizer for d, so the denominator in Eq. 3 is appropriate for inference but, I suggest, does not give a readily interpretable version of Cohen's d if used as a standardizer.

Problems with δ^*

First, consider the base case of $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and $N_1 = N_2$, so that $q_1 = q_2 = ½$. Equation 2 and my discussion above give s_p as our preferred standardizer; in addition, using s_p as our estimate for σ is the most common way to calculate d. However, the denominator in Eq. 3 reduces to 2σ, not σ, and so $\delta^* = \delta / 2$. Therefore, in the equal-variance, equal-group-sizes case, δ^*, as used by Shieh (2013) and Kulinskaya and Staudte (2007), does not reduce to Cohen's δ as most commonly defined.

Second, the standardizer for δ^* (the denominator in Eq. 3) is, as was mentioned above, chosen as relevant for inference and not as necessarily giving a measurement unit for d that has any clear conceptual meaning or any generality over aspects of the experiment—including relative sample sizes and different experimental designs. Its most striking lack of generality is its dependence on the relative sample sizes, indexed by q_1 and q_2: Choose different sample sizes for your experiment, and the measurement unit of δ^* changes.

These two problems with δ^* mean that d^* values are not comparable with familiar d values. In addition, when experiments vary even a little in their characteristics, d^* values are not comparable across experiments and cannot be combined by meta-analysis. For research practice, I suggest that these are fatal flaws.

Shieh (2013) defended δ^* by noting that it is directly related to "the strength of association effect size or weighted coefficient of determination . . . under the heteroscedastic ANOVA models when there are two populations" (p. 3). Perhaps, therefore, δ^* should be regarded as a member of a family of ES measures, including such measures of association, and given a justification and interpretation in terms of its relationship with those measures. It is not my purpose to suggest how that might be done or to defend such an approach. I do suggest, however, that consistency with the most familiar version of Cohen's d is likely to be more important and useful in practice than such a relationship. Also, in further justification of δ^*, after explaining the dependence of δ^* on relative sample sizes, Shieh described another case of an ES measure depending on characteristics of the experiment:

> Note that the squared multiple correlation coefficient is the prevailing strength of association effect size in linear regression. Despite its usefulness, applied researchers may not notice that it is a function of both the model . . . parameters and the distribution properties . . . of the designated covariates. (p. 3)

That another ES measure is dependent in this way does not lessen the disadvantage that δ^* lacks the generality necessary to meet the two requirements for d that I describe in the first paragraph of this comment.

A heteroscedastic example

Imagine that we study IQ scores, using a well-established test. We have two independent groups, one from the general population (assume $\sigma_1 = 15$), the other from a sect we believe has relatively less variable IQ scores (assume $\sigma_2 = 7.5$). We take two samples, obtain two means, and wish to calculate d for the difference. For simplicity, assume that the two sample *SD* values happen to equal the respective population *SD* values. Using $s_1 = 15$ as the standardizer almost certainly makes the best conceptual sense, although perhaps a case could be made for $s_2 = 7.5$ if we have strong reasons for using the population of scores in the sect as our reference. Equation 3, however, uses as standardizer the square root of a weighted average of the two variances, where the variance corresponding to the smaller sample is assigned greater weight. That is appropriate for inference, but I cannot see how it provides a measurement unit for d that researchers or readers could readily grasp or interpret in the applied setting.

Suppose that we find an original-units difference between our two sample means of 5 points on the IQ scale. Whatever the sample sizes, the most meaningful d is $5 / s_1 = 5 / 15 = 0.333$, or, just possibly, $5 / s_2 = 5 / 7.5 = 0.667$. I suggest that it makes the best conceptual sense to evaluate the difference of 5 points against a comparison control *SD* of 15, or perhaps against the sect *SD* of 7.5. This example illustrates the challenge of heteroscedasticity: Whether either of those standardized ES measures—or δ^*— makes sense needs to be considered in the particular context (Kline, 2013, pp. 137–138).

Now consider the three sample size splits used by Shieh, which were 15/15, 10/20, and 20/10, so the q_1 values were, respectively, .5, .333, and .667. For our example, Eq. 3 gives values of d^* that are, respectively, 0.211, 0.181, and 0.222. These values are quite different from the conventional $d = 0.33$ and, furthermore, vary merely because relative sample sizes vary. The variation of d^* with q_1 may seem modest, but manipulating q_1 can easily give much greater variation: Different experiments might give values of d^* that suggest, in one case, a small effect and, in another case, a large effect, even though the original-units effect was the same in each case. More generally, it is difficult, if not impossible, to give a meaningful interpretation in the research context of the measurement unit of d^*, and therefore, d^* values do not help us understand our results, nor are they likely to be usable in a meta-analysis.

This example prompts broader consideration of d in the heteroscedastic case. Suppose that we use my recommended d standardized by s_1, the control group *SD*. Does it make conceptual sense to compare, or combine by meta-analysis, such a d value with d values calculated for situations likely to have homogeneity or different extents of heterogeneity (differing experimental group population variances)? As usual, judgment in context is required. However, suppose that we have a number of studies that assessed the effectiveness of a training procedure intended to increase IQ scores. Each compares an experimental sample from one of a number of special populations with a control sample from the general population. The special populations might have diverse variances, but we might still judge it meaningful to compare or combine the different d values, all standardized to the control *SD*, despite the differing amounts of heterogeneity. (We might use as standardizer an *SD* estimate pooled over all the control samples; Cumming, 2012, p. 289.)

Conclusion

My example above suggests that if we suspect heteroscedasticity, we should, as usual, choose a population *SD* that makes the best conceptual sense as the measurement unit for δ. That is likely to be the *SD* of the control (or reference, or base case) population. Noting the heteroscedasticity, our best estimate of that *SD* will be our sample *SD* for the control group, rather than an estimate pooled across experimental and control groups. Given heteroscedasticity, inference on the difference poses the challenges that Hedges (1981) and Shieh (2013) addressed, but choosing a standardizer for d should be no more difficult than preference for the control group *SD* over a pooled value (Grissom & Kim, 2012, p. 68). I suggest that researchers should prefer such a d to the d^* discussed by Shieh.

Acknowledgments This research was supported by the Australian Research Council.

References

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie / Journal of Psychology, 217,* 15–26. doi:10.1027/0044-3409.217.1.15

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532–574. doi:10.1177/0013164401614002

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist, 60,* 170–180. doi:10.1037/0003-066X.60.2.170. Available from tiny.cc/inferencebyeye.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York: Routledge.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128. doi:10.2307/1164588

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington DC: American Psychological Association.

Kulinskaya, E., & Staudte, R. G. (2007). Confidence intervals for the standardized effect arising in the comparison of two normal populations. *Statistics in Medicine, 26,* 2853–2871. doi:10.1002/sim.2751

Shieh, G. (2013). Confidence intervals and sample size calculations for the standardized mean difference effect size between two normal populations under heteroscedasticity. *Behavior Research Methods*. doi:10.3758/s13428-013-0320-7. Published online: 7 March 2013.